De novo missense mutations in neurodevelopmental disorders

Madeleine R. Geisheker

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Evan E. Eichler, Chair

Gail P. Jarvik

Raphael Bernier

Program Authorized to Offer Degree:

Genome Sciences

University of Washington

Abstract

De novo missense mutations in neurodevelopmental disorders

Madeleine R. Geisheker

Chair of the Supervisory Committee:

Evan E. Eichler

Genome Sciences

Autism spectrum disorder (ASD) is a pervasive neurodevelopmental disorder (NDD) with a high prevalence in the US (1 in 59 children). It is commonly comorbid with other NDDs such as developmental delay (DD), intellectual disability (ID), and epilepsy (EPI). In this thesis, I examine the role of *de novo* missense mutations in NDDs with a goal of identifying genes and specific mutations that are candidates for pathogenicity. I characterize the aggregate signal for *de novo* missense mutations in 8,477 NDD cases, finding both quantitative and qualitative differences between mutations in cases and controls. I also find 40 amino acids that bear *de novo* substitutions in two or more unrelated individuals and develop a tool to assess the likelihood of these observations in the context of stochastic *de novo* events. I then use targeted sequencing to further establish the association of these recurrent mutations with disease. Upon finding the same p.Ala646Thr substitution in five cases in glutamate receptor subunit GRIA1, I carry out functional experiments that show alterations in ion flux. I also assessed clustering of *de novo* missense mutations as this pattern is associated with NDDs, such as Schinzel-Giedion syndrome. I used an unsupervised clustering algorithm, CLUMP, to compare the distribution of *de novo* missense mutations in NDD cases with private missense events in controls and found 200 genes that were significantly more clustered ($p < 0.05$). As this set of genes is enriched for neuronal functions, a known association of NDD risk genes, it is likely that clustering is a valid feature for identification of disease genes. With increased

exome sequencing on NDD cases, I was able to assess *de novo* mutation burden in 10,927 cases with ASD, DD, or ID. With two different models, I found 253 total genes with more *de novo* mutations than expected, 123 of which have a burden of missense mutations. Protein-protein interaction and enrichment analyses of genes with a burden of mutation finds that those with a burden of truncating mutations have roles in transcription regulation while those with missense burden have roles in synaptic signaling. This same neuronal enrichment, including in the amygdala and cortex during fetal development, is seen in genes with clustered *de novo* missense mutations. Interestingly, the phenotypes of patients with missense mutations in a novel gene, *TRRAP*, segregate with mutation clustering, suggesting the biological relevance of this pattern of mutation. As burden analysis only identified some of the expected pathogenic NDD genes, I included mutations from patients with EPI to my discovery set. Novel genes identified with this addition are enriched for expression in the striatum. Targeted sequencing of these hotspots of mutation identified additional substitutions at 20 recurrent sites and established 28 new recurrent sites. Eighteen of the sites are known to be pathogenic, and some evidence supports the disease association of the remaining 30 sites. Continued assessment of genes with these patterns of mutation, as well as expansion into gene families, will help to characterize the genetic architecture of NDDs, specifically missense mutations, and provide increased understanding of brain development and pathogenesis.

Table of Contents

List of Figures

List of Tables

Acknowledgements

This work would not have been possible without the help from so many other people. First and foremost, I owe many thanks to my mentor, Evan Eichler. His outstanding career and desire to answer important questions about human biology enabled him to orchestrate an international network of clinicians and researchers, and through this collaboration, we are able to study thousands of individuals with neurodevelopmental disorders. I am so grateful to all of the families who contribute to this work, as they are truly the source of our data. This work would not be possible without them.

I would also like to thank those who contributed to my development as a scientist and inspired me to pursue research. First, Joan Stiles, whose brilliant talk on brain development in children with perinatal strokes sparked my interest in research. I owe many thanks to Kathy French and Bill Kristan, who offered me a position in their lab, and despite my lack of experience, allowed me to work on a project independently. My experience in their lab fostered my appreciation of scientific inquiry and confirmed my interest in research. I am also so grateful for the mentoring that Frank Haist provided, helping me to design a project and supporting me through all of the challenges. It was my work with him that ignited my desire to pursue my PhD.

I owe so many thanks to my friends who provided such a strong support system. That there are far too many of them to name is testament to the inclusiveness of our group. At the same time, the connections we have with each other are so strong that they feel like family, and we share so much love and so many laughs. It still amazes me that I met such an amazing group of people, and I feel so lucky to have them in my life. Through them, I have met so many more wonderful people, including my other half. Brendan, you are such an amazing human and I am so happy to have you in my life. You inspire me to ask more questions, to be a better person, and to spread more love.

Finally, I would like to thank my family. I moved to Seattle to be closer to extended family, and I am so glad I made that decision. The peace of mind that comes from being surrounded by loving people who will help me in any situation is so wonderful, especially with the stresses that come with graduate school. As

with my friends, I can always turn to them for love and laughs. Most importantly are my parents, who raised me to be curious and independent, and who support me always. I will never be able to put into words how grateful I am for everything they have done for me. Mom and Dad, I love you so much!

Chapter 1. Introduction

Autism spectrum disorder (ASD) is a pervasive neurodevelopmental disorder (NDD) characterized by social impairments and restricted behavior and interests. Early observations of ASD in patients with other single-gene disorders prompted research into its genetic etiologies. As technology has developed, findings have established a causative role for *de novo* copy number variants (CNVs) and single-nucleotide variants (SNVs) in sporadic ASD. Despite increasing cohort sizes, such as with the >2,500 families with a child with ASD in the Simons Simplex Collection, the number of pathogenic genes identified has not reached its expected maximum, especially for missense mutations. To improve identification of genes with likely gene-disruptive (LGD) mutations (nonsense, frameshift, or splice site), cohort sizes were further increased by studying multiple NDDs, as they are extensively comorbid and have shared genetic etiologies. Additionally, targeted sequencing methods were developed to efficiently and affordably capture sequence at loci of interest. Despite these large increases in sample size, the number of genes with a burden of missense mutations remains far below the expectation. As *de novo* missense mutations are expected to account for more cases of ASD than LGD mutations, methods to differentiate potentially pathogenic missense events from the high rate of incidental events are needed. The goals of this thesis are to characterize the missense signal in NDDs, to identify novel genes and mutations that are pathogenic in NDDs, and to use phenotypic and functional evidence to strengthen the association of specific genetic events with NDDs.

1.1 Autism spectrum disorder

Autism spectrum disorder (ASD) is defined by persistent deficits in social communication and interactions and restricted, repetitive patterns of behavior or interests[1]. In the previous version of the Diagnostic and Statistical Manual of Mental Disorders[2], social communication and social interactions were separate diagnostic categories. While children with no language or cognitive delays were previously given a separate diagnosis of Asperger's disorder, they are now included in the broader ASD diagnosis. ASD is increasingly common, with current estimates of 1 in 59 children diagnosed in the United States[3]. Prior to the late 1970s, the cause of ASD was not thought to be biological but instead environmental; parenting styles were frequently vilified. When the co-occurrence of ASD and known single-gene disorders such as

fragile X syndrome was noticed in the 1980s, the genetic nature of ASD became apparent[4]. The most recent estimates of heritability, based on twin and sibling studies, place it at 83%[5]. ASD's high prevalence and heritability have driven research towards discovery of causal genetic events that might explain pathogenesis and lead to improved treatment methods.

1.2 Early findings on the genetics of ASD

The search for the genetic causes of ASD distinct from single-gene disorders has advanced with technology. In the 1990s, studies on families in which multiple children have ASD were carried out in an attempt to link specific genes with the disorder. Early searches focused on genes that were known causes of other disorders such as *FMR1* in fragile X[6]. Later, scientists widened their search to the whole genome using linkage analyses and found associations between ASD and regions on eight chromosomes[7], but they were unable to unambiguously identify specific causative genes. While these studies were ongoing, a new technology was developed to compare genomic imbalances in tumor cells[8]. This technique, called comparative genomic hybridization (CGH), enabled identification of deletions and duplications of segments of DNA. These CNVs alter gene dosage, and duplications can further disrupt genes. The capabilities of CGH were further advanced by combining it with DNA microarray techniques, which enabled detection of CNVs at a resolution of 100 kilobases and comparison of a patient's genome with a reference genome[9]. With the higher-resolution array CGH method, several groups identified large (>50 kilobases) events in children with idiopathic, i.e., without a known cause, intellectual disability (ID) and congenital anomalies[10–14]. Importantly, some of these events were found to be *de novo*, and children with the same CNV had similar phenotypes. These results were strong support of the causative nature of CNVs in sporadic syndromes with NDD features.

1.3 *De novo* CNVs in ASD

De novo events are new germline mutations that are not inherited from either parent. They have long been known to cause sporadic genetic disease[15]. Rett syndrome, for instance, in which ASD is a common phenotypic feature, was found to be caused by a *de novo* event on the X chromosome[16]. However, technological limitations slowed the discovery of pathogenic *de novo* events. With the advent of array

CGH, sporadic disease could be studied using high-throughput approaches. The discovery of *de novo* CNVs in ID prompted a search for these events in ASD, another disorder that is commonly sporadic. In 2007, Sebat et al. published the first study assessing *de novo* CNVs in children with idiopathic ASD[17]. They searched for CNVs in 165 families with one or more children with ASD, excluding any with known genetic events and families where the affected child also had severe ID or congenital anomalies. Of these families, 118 were simplex, meaning only one child was affected, and 47 were multiplex, with multiple affected children. For comparison, they assessed 99 control families in which no members had ASD. They found a total of 17 *de novo* CNVs in 14 children with ASD and two in children from control families. The rate of *de novo* CNVs in ASD cases was significantly higher than in the rate in controls ($p = 0.0005$), suggesting the role of gene dosage alterations in ASD. Further, this association was driven by simplex cases ($p = 0.0005$), as the rate of *de novo* CNVs in multiplex cases was not significant ($p = 0.59$). This finding is consistent with the causality of *de novo* CNVs in simplex ASD, resolving the high heritability of ASD with its sporadic occurrence. However, the implication of CNVs as casual in ASD did not identify any specific genes as responsible for pathogenesis. Discovery of these genes is critical for improving patient outcomes because it provides mechanistic information that supports the development of treatments that target the causes of ASD.

1.4 *De novo* SNVs in ASD

Armed with the knowledge that *de novo* events explain some cases of sporadic ASD and with increasingly affordable sequencing technologies, O'Roak et al. looked for SNVs in 20 simplex families[18]. The authors used next-generation technologies for high-throughput exome sequencing, as mutations in coding regions are more likely to be disruptive. They sequenced both affected children and their parents for identification of *de novo* events. In these 20 families, they identified 21 *de novo* mutations, including one frameshift mutation, one splice site mutation, and nine missense mutations predicted to be damaging by PolyPhen-2. These 11 protein-altering amino acid changes occurred at highly conserved positions in the genome. Importantly, several of the mutations occurred in genes that had been previously associated with other NDDs and genes known to be critical for brain function, such as *GRIN2B*, a subunit of glutamate receptors critical for learning and memory[19,20]. Although each of the mutations occurred in a

different gene, preventing locus-specific statistical analysis, this provided strong evidence for the causal role of *de novo* SNVs, and therefore disruptions in single genes, in ASD pathogenesis. However, discovery of a single mutation in a gene, even if the gene has a known role in neurobiology, is insufficient to ascribe pathogenicity to that gene. Many of these individuals also had a second or third *de novo* event in a different gene, further complicating assessment of causality. A much larger number of people with ASD would need to be studied to see enough mutations in a gene to confidently implicate it in disease.

1.5 Gene identification with increased cohort sizes

To address this problem, the Simons Foundation Autism Research Initiative (SFARI) was formed[21]. Members of this group recruited over 2,500 simplex families for exome sequencing as in O'Roak et al. (2011). Importantly, in addition to requiring that each family had only one child with ASD, they also recruited families with at least one unaffected child as this increases the likelihood that the etiology of the affected child is *de novo* mutation. Further, unaffected siblings are ideal genetic controls[21]. Not only did SFARI perform sequencing on each child and parents, but they also gathered detailed and standardized phenotype information on the whole family. For the children with ASD, they conducted many different neuropsychological exams to assess several domains of neurodevelopment, such as intelligence and communication. They also gathered information on other common comorbidities and medical problems as these could also be explained by a genetic event and further link the mutation to disease via biological mechanisms. Additionally, they assessed parents and unaffected children for ASD characteristics, intelligence, and other factors to further delineate the aspects of each child's phenotype that are inherited as opposed to caused by a *de novo* event.

Several papers were published on this cohort as it grew[18,22–26]. By looking at estimated *de novo* rates, a significant excess of mutations was seen in specific genes[25,27], including *NTNG1*, *CHD8*, and *SCN2A*. However, predictions at this time suggested that between 384 and 821 loci were involved in pathogenesis, indicating that many more patients needed to be tested[25]. In 2014, a paper on the full cohort of 2,517 families was published[28]. With increased sample size and the inclusion of unaffected siblings, the authors could better compare rates of *de novo* SNVs. The rate of synonymous *de novo*

mutations was not significantly different between affected and unaffected children (0.34 and 0.33 events per child, respectively) (**Figure 1.1**). In 2,508 affected children who were successfully sequenced, they found 391 LGD mutations. The rate of LGD mutations in affected children is 0.21 events per child; the rate in unaffected children is 0.12. As these rates are significantly different ($p = 2x10^{-5}$), there is a strong signal for *de novo* LGD mutations in ASD. The excess of mutations per child in affected children (0.21 - 0.12 = 0.09 events per child) can be attributed to ASD. Importantly, 27 genes were hit by two or more LGD mutations and most of these recurrently hit genes cluster into four categories: fragile X protein targets, chromatin modifiers, proteins in the postsynaptic density, and proteins that are embryonically expressed. All of these categories are biologically plausible in ASD pathogenesis.

This paper was also the first to establish a significant aggregate signal for missense mutations, with a rate of 0.94 events per affected child and 0.82 events per unaffected child. Previous papers on smaller subsets of the cohort did not find this signal, likely because the sample size was too small[25,26]. By the same calculation as for LGD mutations, 0.12 missense events per child are attributable to ASD. Those authors predict that the fraction of missense events seen in children with ASD that are likely to be pathogenic is only 13% (0.12/0.94). This low rate is due to the inherent variability in effect of missense mutations and their high frequency. While some missense mutations can be as damaging as an LGD mutation based on the amino acid change or location of the mutation, some have limited functional effect.

The *de novo* missense signal established by that paper was significant in multiple ways. First, it illustrated that many more people with ASD will need to be studied to find all of the contributing genes, as a cohort of 2,508 was insufficient to significantly associate any gene with ASD given the background rate of nonpathogenic *de novo* mutations. Second, it explained some of the missing heritability in ASD. Only 9% of cases of ASD can be explained by mutations in genes identified due to LGD mutation burden, but family studies estimate that 83% of cases have genetic causes[5]. As missense mutations occur nearly eightfold more frequently than LGD mutations, they are expected to be implicated in 12% of cases of ASD. Despite this, Iossifov et al. did not explore the distribution of these events. Part of this may have been due to the limited number of likely damaging missense mutations. Since only 13% of those seen

(0.12/0.94) are likely to be attributable and few methods were available to predict which ones are, many more damaging mutations are needed to see recurrence in a gene and establish gene-specific burden. With identification of new genes through increasing cohort sizes, research on specific mechanisms of pathogenesis will be possible, enabling improved and specific treatments and a better understanding of the complexities of brain development.

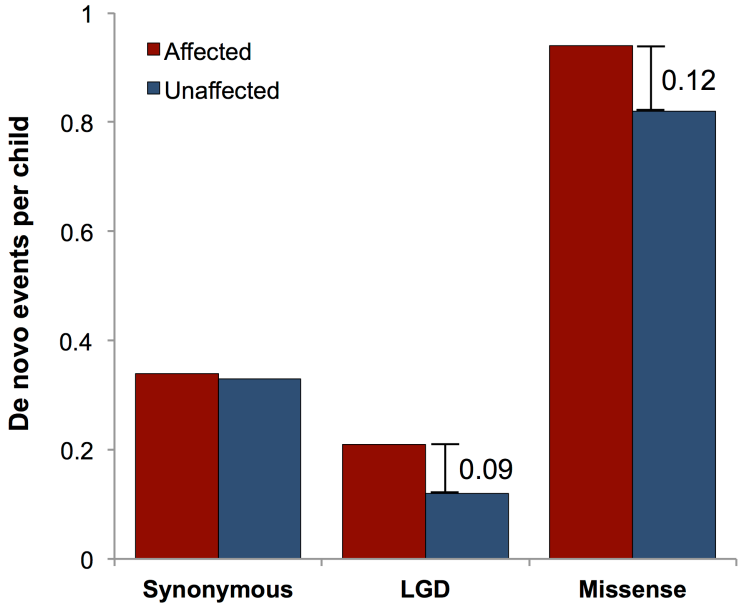

Figure 1.1. Rate of de novo SNVs in children with ASD and their unaffected siblings. Data from Iossifov et al. (2014). The rates of synonymous mutation in affected (0.34) and unaffected (0.33) children are not significantly different. The rate of *de novo* LGD mutations in affected children is 0.21 and the rate in unaffected children is 0.12. For *de novo* missense mutations, the rates are 0.94 (affected) and 0.82 (unaffected). As the events in unaffected children are assumed to be nonpathogenic, the ascertainment differential between the two groups can be attributed to pathogenic events. Although unaffected children have a high rate of *de novo* missense mutations, the differential for these mutations is higher (0.12) than the differential for LGD mutations (0.09).

1.6 Comorbidities in NDDs

With a goal of identifying genes with *de novo* missense burden, we can take advantage of the phenotypic and genotypic commonalities of NDDs to increase our sample size. Nearly one-third (31%) of individuals with ASD also have ID (IQ < 70), and 25% have borderline ID[3] (full-scale IQ [FSIQ] 71-85). Additionally, 22% of individuals with ASD also have epilepsy[29,30]. For individuals with ID, 28% were also diagnosed with ASD[31] and 26% also have epilepsy[32]. Of patients with epilepsy, 80% have an FSIQ under 85, and 40% have ID[33]; 21% of individuals with epilepsy also have ASD[34]. Importantly, in children with epilepsy, a

strong risk factor for comorbid ASD is IQ, wherein the prevalence of ASD is 2.2% in children with an IQ > 80 and 13.8% in those with an IQ < 80[35]. Additionally, 3.7%–5.2% of children with ID also have schizophrenia[36], and genes implicated in schizophrenia overlap with genes implicated in other NDDs[37]. One example is the 15q11.3 microdeletion, which can cause ASD, ID, epilepsy, and schizophrenia in different patients[38]. The high rate of comorbidities means that most studies are composed of patients with multiple NDD diagnoses, although the primary focus may be only one.

1.7 Targeted sequencing of candidate loci

As described in the previous sections, discovery of genetic etiologies of ASD and other NDDs has advanced with technology. The development of array CGH enabled discovery of *de novo* CNVs, and the development of high-throughput exome sequencing enabled discovery of *de novo* SNVs. However, even after sequencing over 900 individuals with ASD, few genes had recurrent *de novo* mutations[39]. Instead of using whole-exome sequencing on additional cases, a targeted sequencing approach was used to assay a subset of genes[39–42]. This method uses molecular inversion probes (MIPs) to capture loci of interest in a highly multiplexed manner and for a cost of less than $1 per gene per sample. A later development to the method was the addition of single-molecule tags (smMIPs), which enables consensus calling[43]. This reduction in per-sample cost relative to whole-exome sequencing allows for a much larger sample size at potential risk genes. This is critical for ASD and other NDDs, which are genetically heterogeneous and often caused by *de novo* mutations that are individually rare.

Using smMIPs targeting 64 candidate genes, O'Roak et al. (2014) identified 56 rare *de novo* mutations in 2,527 probands[44]. These mutations occurred in 27 of the targeted genes, and several genes were mutated recurrently. To test the association between genes with recurrent mutations and disease, the authors developed a model to predict the number of *de novo* mutations expected for a gene. With this model, nine genes had a burden of total *de novo* mutations. While this was progress, it is far from the predicted hundreds of involved genes and further increases in the number of cases studied are needed.

Another method to increase sample size is to study patients with a variety of NDDs, as they are commonly comorbid and have shared genetic etiologies. By including studies with different primary focuses, we can effectively improve our ability to detect risk genes. This comes at a cost of reduced disease specificity, but later studies on larger, more specific cohorts may be able to differentiate disease-specific genes and mutations. Lack of strict phenotypic distinction complements the genotype-first approach, wherein patients are grouped based on common genotypes[45]. This approach is especially successful for NDDs[46], where each genetic event is individually rare and phenotypes are often multisystemic and variably expressed. By increasing sample size in two ways, assessing broad NDDs jointly, and using targeted sequencing, we hope to detect a gene-specific signal for *de novo* missense mutations that we can further pursue to establish potentially pathogenic genes and better understand these complex disorders.

1.8 Thesis goals

This introduction has described the high heritability of ASD, and particularly the role of *de novo* genetic events in simplex ASD. While missense mutations are expected to make an important contribution, their variable effect and high incidental rate have limited discovery of genes that cause pathogenesis in this manner. The overarching goal of this thesis is to identify such genes and the features that define them, as this will enable improved understanding of typical and atypical brain development and provide a path for future therapeutics. We will accomplish this in three ways:

1) **Characterize *de novo* missense signal in NDDs.** In Chapter 2, we study *de novo* missense mutations in cases with NDDs and show both quantitative and qualitative differences in events seen in cases versus events seen in controls. We use a model to identify 35 genes with a burden of *de novo* missense mutations. We also identify sites, i.e., amino acids that bear *de novo* substitutions in two or more unrelated individuals, some of which have been previously associated with NDDs. Additionally, in multiple sets of genes identified with different methods throughout this thesis, we show enrichments for genes with synaptic function and expression in neurons during fetal development.

2) **Identify novel genes and mutations that may contribute to disease.** As available models identify fewer genes with *de novo* missense burden than expected, we develop a different model that takes into account the importance of missense mutation location and identify sites with burden. To further increase discovery of potential risk genes, we use a clustering algorithm to identify genes with mutations in closer linear proximity to each other in cases than in controls. Both of these phenomena, sites and clusters, have been associated with disease and, in this thesis, findings support their validity as features of risk genes. In Chapters 3 and 5, we use targeted sequencing to vastly increase sample size in regions with these patterns of mutation and identify additional events that may be associated with disease.

3) **Use phenotypic and functional analyses to build strength of association between *de novo* mutations and NDDs.** In Chapter 2, we identify two individuals with the same *de novo* missense mutation in *GRIA1*, a subunit of AMPA glutamate receptors. Targeted sequencing in Chapter 3 identifies three additional unrelated individuals with the same mutation. This receptor is critical in learning and memory formation, and evidence from mice shows that this mutation affects neuronal ion channels and leads to a strong phenotype. We performed functional studies and showed, for the first time, that this event in humans has similar effects. Furthermore, individuals with this mutation share phenotypic features. Assessment of the gene *TRRAP* with a burden of *de novo* missense mutations and significant clustering, also shows a correlation between genotype and phenotype. Notably, severity of disease in these individuals segregates with the mutation clustering pattern. These findings support the role of these *de novo* missense mutations in disease, and suggest that additional mutations warrant follow-up for improved understanding of disease.

Chapter 2: Patterns of *de novo* missense mutations in NDDs

This chapter has been adapted from: Geisheker, M.R., Heymann, G., Wang, T., Coe, B.P., Turner, T.N., Stessman, H.A.F., et al. (2017). Hotspots of missense mutation identify neurodevelopmental disorder genes and functional domains. *Nature Neuroscience*, 20(8):1043-1051. doi:10.1038/nn.4589.

2.1 Summary

Although *de novo* missense mutations have been predicted to account for more cases of autism than LGD mutations, most research has focused on the latter. We initially defined the properties of potentially pathogenic *de novo* missense mutations by studying the distribution pattern of such mutations from published exomes in individuals with more broadly defined NDDs including ASD. We find a significant increase in both the rate and predicted deleteriousness of *de novo* missense mutations in cases compared to controls. Thirty-five genes have a significant burden of *de novo* missense mutations, including several that have been associated with NDDs and 14 where no LGD mutations have been observed in patients. Genes with an excess of missense mutations are significantly enriched for synaptic functions, including FMRP targets but not targets of CHD8 binding. Mutations in genes with two or more *de novo* missense events were more likely to occur in cases and were more severe. Notably, 40 amino acid sites in 36 genes have *de novo* missense mutations in multiple unrelated cases.

2.2 Introduction

Multiple lines of evidence provide strong support for a genetic basis for ASD. *De novo* mutations, originating primarily in the parental germline, are individually rare but their collective risk is substantial and accounts for an estimated 30% of simplex ASD cases[28,47]. To date, most of the emphasis on identifying high-impact risk variants has focused on establishing burden for LGD mutations (nonsense, frameshift, or splice-site)[45,48,49]. High-impact risk genes with primarily *de novo* missense mutations have been understudied because a much smaller fraction (13%) are thought to be pathogenic when compared to *de novo* LGD mutations (42%)[28]. Moreover, *de novo* missense mutations are eightfold more common making

it more challenging to prove their statistical relevance. Notwithstanding, a comparison of mutation rates in individuals with ASD and their unaffected siblings reveals that *de novo* missense mutations contribute to disease risk in as many, if not more, cases than LGD mutations (12% vs. 9%, respectively)[28].

The identification of genes with a significant burden of missense mutations, then, is likely to highlight new classes of neurodevelopmental disorder (NDD) risk genes. In some cases, this may reflect genes with such critical functions that LGD mutations are incompatible with life[28,50]. In other cases, the mutation's effect on the protein may differ. For example, missense mutations are more likely to have a gain-of-function effect[51] when compared to LGD mutations, which are predominantly loss-of-function. Clustering of missense mutations may highlight important and even novel functional domains, providing insight into ASD pathogenesis and future downstream therapeutic targets. High-confidence ASD risk genes have been successfully identified by searching for mutation recurrence[45,48,52,53]. Given that missense mutations are more common and ~90% of them are thought to be incidental[28], i.e. not pathogenic, a much larger sample size is required to prove pathogenicity. We took advantage of the significant phenotypic and genotypic overlap between ASD, DD and ID, EPI, congenital heart disease (CHD), and schizophrenia[54] (SCZ) to study the pattern and distribution of *de novo* missense mutations more broadly.

2.3 Methods

Exome datasets and missense mutation annotation. We initially analyzed all *de novo* missense mutations available from 24 published cohorts[28,39,44,55–78] of *de novo* mutations in individuals with NDDs (denovo-db v.0.9; **Appendix 2.1**)[79]. The NDD set included 8,477 individuals diagnosed with ASD, DD, ID, EPI, CHD, and SCZ, as well as four cohorts of unaffected controls[28,55,80] (N = 2,178) (**Table 2.1**). Only CHD patients from Homsy et al. (2015) with a secondary diagnosis of NDD were included in this study; we also excluded unaffected siblings of ASD patients as controls if they had a Social Responsiveness Scale (SRS) score ≥60 to remove controls on the autism spectrum[81]. Variants were annotated with SeattleSeq[82] version 138, which provides annotation for all available RefSeq transcripts in GRCh37/hg19. In the case of multiple transcripts, we selected the transcript for which the majority of missense mutations were annotated in both cases and controls. All *de novo* missense mutations were either previously

validated or investigators relied on a high (>95%) validation rate in a subset of mutations to ensure specificity. As some individuals with ASD were assayed as part of multiple cohorts, we took care to remove any duplicate entries. When possible, we compared the global identifier given to the samples that were housed at Rutgers (RUID). Three duplicate entries were found in this manner. For other shared mutations in ASD cohorts, we performed PCR amplification and Sanger sequencing on in-house DNA samples to confirm secondary variants. Five out of six pairs tested (two ASC [Autism Sequencing Consortium]-SSC [Simons Simplex Collection] pairs and three ASC-TASC [The Autism Simplex Collection] pairs) shared a second variant and we therefore assumed them to be duplicates. The presence of uniquely identifying secondary site mutations was also used to eliminate potential duplicates for globally dispersed samples. We excluded high-frequency mutations (MAF > 0.1%) observed in NHLBI GO ESP Exome Variant Server (Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP), Seattle, WA (http://evs.gs.washington.edu/EVS/) [August 2016]).

Table 2.1. Discovery cohorts in denovo-db v.0.9.

Study diagnosis	Cohorts	Individuals	Trios*	Quads**	*De novo* missense mutations†	*De novo* LGD mutations	Studies
Autism spectrum disorders (ASD)	8	4197	2120	1956	3185	677	De Rubeis 2014[75] (ASD1), Hashimoto 2015[56] (ASD2), Jiang 2013[59] (ASD3), Lee 2014[62] (ASD4), Michaelson 2012[65] (ASD5), Simons Simplex Collection[25,28,39,61] (ASD6), Tavassoli 2014[67] (ASD7), Yuen 2015[70] (ASD8)
Congenital heart disease (CHD)	2	775	775	0	308	152	Homsy 2015[58] (CHD1), Zaidi 2013[71] (CHD2)
Developmental delay (DD)	4	2104	2104	0	1545	486	de Ligt 2012[74] (DD1), Lelieveld 2016[63] (DD2), Hurles 2014[83] (DD3), Rauch 2012[66] (DD4)
Epilepsy (EPI)	6	602	601	1	267	76	Barcia 2012[73] (EPI1), Dimassi 2015[77] (EPI2), epi4k 2013[84] (EPI3), Helbig 2016[57] (EPI4), Veeramah 2012[68] (EPI5), Veeramah 2013[69] (EPI6)
Schizophrenia (SCZ)	4	799	715	84	502	87	Fromer 2014[78] (SCZ1), Gulsuner 2013[55] (SCZ2), Kranz 2015[60] (SCZ3), McCarthy 2014[64] (SCZ4)
TOTAL CASES	24	8477	7115	2041	5807	1478	
Unaffected	4	2178	270	1908	1475	237	GoNL 2014[80], Gulsuner 2013[55], Rauch 2012[66], Simons Simplex Collection[28]

*Family with unaffected parents and one affected child

**Family with unaffected parents, one affected child, and at least one unaffected child

†Missense mutations with minor allele frequency (MAF) < 0.1% in ESP (N = 6,503)

Statistical analyses. Wherever possible, nonparametric tests were used. Data collection and analysis were not performed blind to the conditions of the experiments. Burden was compared between cases and controls for rare (MAF < 0.1% in ESP) *de novo* missense mutations. Comparisons in rate of mutation and

gene recurrence were made using two-sided Fisher's exact tests. For comparisons of mutation rate and recurrence that depended on identical numbers of cases and controls, we performed one million downsamplings and used permutation tests, reporting the empirical p-values. Data distribution was assumed to be normal but was not formally tested. To identify significant enrichments for missense mutations within genes and genic regions, we applied a probabilistic model that incorporates sequence context and human–chimpanzee fixed differences to generate a null model for the distribution of missense variation across the genome and applied a one-tailed binomial test to test for enrichment[39]. For examination of individual codons and specific target regions, we applied the same method but restricted to the sequence context of the target region and normalized by the gene-specific human–chimpanzee divergence. For all tests we assumed a mutation rate of 1.8 *de novo* coding variants per generation[85]. Multiple testing corrections were applied using two paradigms based on the analysis type. For significance calculations of whole genes, we utilized the Benjamini-Hochberg FDR correction based on an estimated 19,000 genes in the human genome[86] and report the q-values for each test. For codon analysis we applied the conservative Bonferroni family-wise error rate (FWER) correction based on the number of amino acids in the genome ($N = 1.1 \times 10^7$) to generate genome-wide significance estimates and report the adjusted p-value (p_{adj}). Gene ontology enrichment was assessed using PANTHER (database 2017-04-13) for GO biological process annotation and corrected for multiple testing (Bonferroni, reported as p_{adj}). We also applied a one-sided Fisher's exact test for testing the enrichment of specific gene sets, including neuronal compartments such as the post-synaptic density[87] and targets of CHD8[88] and FMRP in brain tissue[89].

2.4 Results

Properties of *de novo* missense mutations in NDD patients. We began by assessing the rates of *de novo* missense mutation in cases and controls. We identified a total of 5,807 *de novo* missense mutations in cases (N = 8,477) and 1,475 such events in controls (N = 2,178) (**Table 2.1**). The fraction of probands with one or more event (50.7%) is significantly greater than the fraction of controls (47.8%; p = 0.016, OR = 1.12 [1.02-1.24], two-sided Fisher's exact test) (**Figure 2.1a**). As there were over three times as many cases as controls, we sought to limit the possibility that the signal is driven by rare outliers in cases and

thus applied a secondary test, downsampling cases to match the number of controls. This further confirmed a significant increase in the rate of *de novo* missense mutations in cases (one-tailed empirical $p = 9.22 \times 10^{-4}$, OR = 1.12 [1.06-1.19], 1×10^6 permutations) (**Figure 2.1a**). While the odds ratios for these two tests are nearly identical, the Fisher's exact test is considered more conservative and the hypergeometric distribution generates a wider confidence bound for the odds ratio when compared to that obtained by simulation.

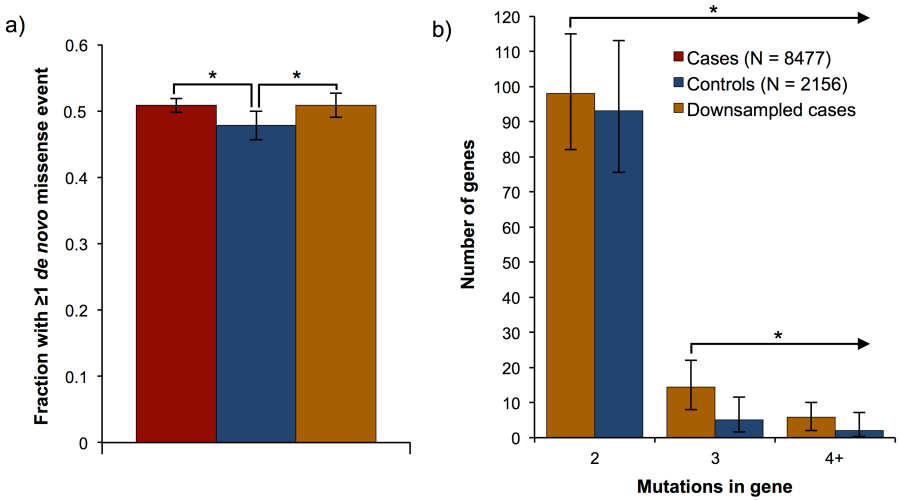

Figure 2.1. Burden and recurrence of de novo missense mutations. Bars for cases and controls represent observed data and error bars indicate the 95% confidence interval (CI) for the observed proportions (Clopper-Pearson method). Box-and-whisker plots for downsampled cases represent the distribution of one million permutations. Boxes show interquartile range (IQR) with lines at the median and whiskers are 1.5 times the IQR. Asterisks indicate $p < 0.05$. **a)** 4,301 out of 8,477 cases (50.7%) and 1,042 out of 2,178 controls (47.8%) have one or more *de novo* missense mutations (denovo-db v.0.9) that are rare in the general population (MAF < 0.1% in ESP). The fraction of individuals with one or more *de novo* missense mutation is significantly higher in cases ($p = 0.016$, OR = 1.12, two-sided Fisher's exact test) even after downsampling (empirical $p = 9.22 \times 10^{-4}$, OR = 1.12 [1.06-1.19]). **b)** The number of genes with two or more mutations in downsampled cases is significantly greater than controls (empirical $p = 0.011$, OR = 1.26 [1.10-1.42]), as is the number of genes with three or more mutations (empirical $p = 3.1 \times 10^{-5}$, OR = 3.13 [2.22-4.03]).

Out of 4,227 genes with rare *de novo* missense mutations in cases, 974 (23.0%) harbor mutations in two or more unrelated cases (**Appendix 2.1**). In contrast, among controls, 101 out of 1,362 genes (7.4%) are mutated recurrently (**Appendix 2.2**). Matching the number of cases and controls, we observe a significant increase in the number of genes among cases with two or more (one-tailed empirical $p = 0.011$, OR = 1.26 [1.10-1.42], 1×10^6 permutations) and three or more (one-tailed empirical $p = 3.10 \times 10^{-5}$, OR = 3.13 [2.22-4.03], 1×10^6 permutations) *de novo* missense mutations (**Figure 2.1b**). The increased recurrence

rate is not explained by increased mRNA or protein length, as genes with recurrent mutations in cases are significantly shorter than those with recurrence in controls (mRNA, $p = 5.19x10^{-3}$; protein, $p = 1.47x10^{-3}$; two-sided Wilcoxon rank-sum tests). Additionally, the total number of genes with mutations is smaller among cases (1,323 in downsampled cases vs. 1,362 in controls), suggesting that mutations in cases are not randomly distributed but rather cluster within fewer genes.

We next compared the severity of *de novo* missense mutations between cases and controls by assessing the Combined Annotation Dependent Depletion (CADD) score[90]. The CADD score distribution is significantly positively skewed in cases compared to controls consistent with an increase in deleteriousness ($p = 2.2x10^{-4}$, two-sided Wilcoxon rank-sum test). Further, at increasing minimum CADD score thresholds, the likelihood that an observed event can be attributed to a case increases (**Figure 2.2a**). At a CADD threshold of 28, the likelihood rises dramatically (>1.2 positive likelihood ratio). Importantly, mutations in genes with higher levels of recurrence in cases also show significantly higher CADD scores ($p = 5.87x10^{-29}$, $F = 45.12$, 3 degrees of freedom, one-way ANOVA), indicating that recurrence and severity are both valuable markers of missense pathogenicity and that they are highly correlated (**Figure 2.2b**).

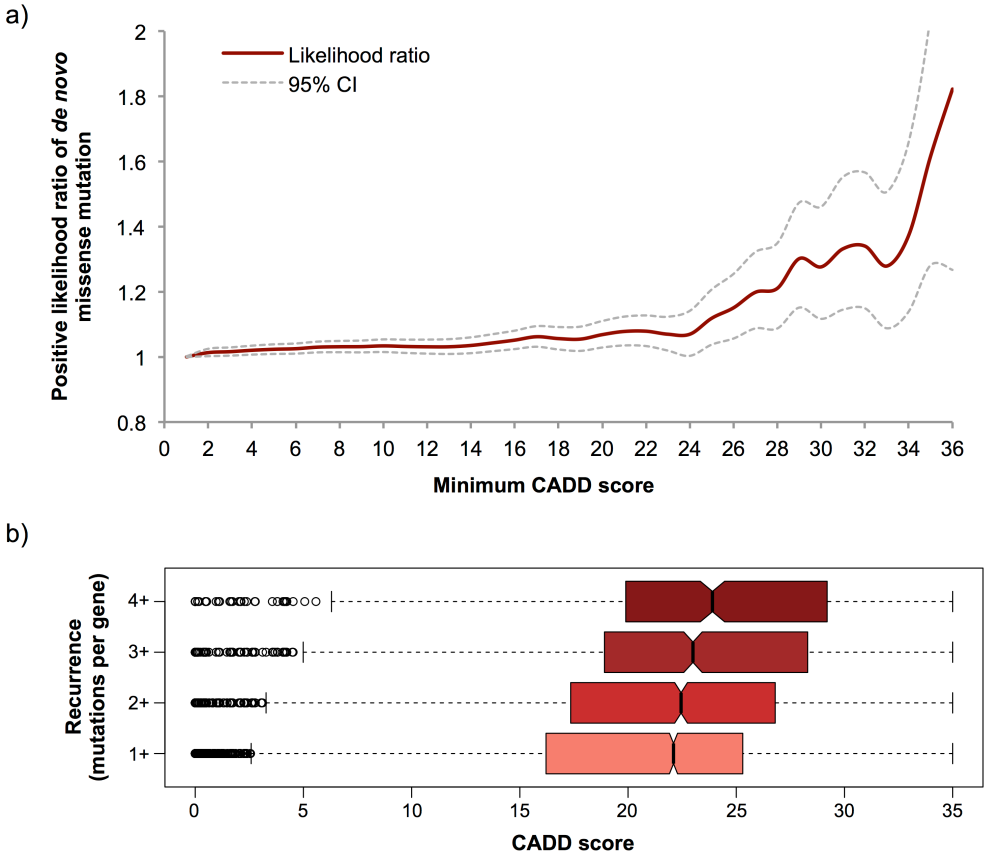

Figure 2.2. Severity of de novo missense mutations. a) *De novo* missense mutations are more likely to be deleterious in cases (N = 5,807 mutations) versus controls (N = 1,475 mutations) and the positive likelihood ratio increases as severity increases (as measured by CADD score). **b)** The distribution of CADD scores skews significantly as the number of *de novo* missense mutations per gene in cases increases ($p = 5.87 \times 10^{-29}$, one-way ANOVA) indicating an enrichment for genes with pathogenic mutations. Boxes show IQR with notches representing the 95% CI of the median; whiskers are 1.5 times the IQR. Circles are outliers.

Genes with recurrent missense mutations. To further assess gene-specific recurrent mutations, we applied a probabilistic model that calculates the expected number of mutations in a gene, based on locus- and base-specific relative substitution rates[39,85]. We identified 35 genes that had significantly more *de novo* missense mutations in cases than expected (false discovery rate (FDR) < 5%) (**Appendix 2.1**). Only two genes, *YIF1A* and *PHKA2*, reached significance in controls (**Appendix 2.2**). For 17 of the genes significant in cases, an excess of loss-of-function mutation has already been established by copy number variants (CNVs) and LGD mutation (e.g., *GRIN2B, PTEN*, and *SCN2A*[46,75,91]). For 13 of the remaining significant genes, no LGD mutations have been identified in the 24 cohorts studied here or in individuals with NDD in the Online Mendelian Inheritance in Man [OMIM; http://omim.org/] or ClinVar

[https://www.ncbi.nlm.nih.gov/clinvar/] databases. While six of these missense-only genes are well known and associated with specific phenotypes (e.g., *PACS1* and Schuurs-Hoeijmakers syndrome[92]), the remaining seven warrant additional follow-up.

As a set, the 35 genes with excess *de novo* missense mutations are enriched for aspects of neuronal communication such as postsynaptic membrane potential regulation (6 observed vs. 0.17 expected, 35.3-fold enrichment, p_{adj} (Bonferroni corrected) = 1.61×10^{-4}, two-sided binomial test) and synaptic signaling (8 observed vs. 0.7 expected, 11.4-fold enrichment, $p_{adj} = 3.30 \times 10^{-3}$, two-sided binomial test), nervous system development (16 observed vs. 3.7 expected, 4.4-fold enrichment, $p_{adj} = 1.05 \times 10^{-3}$, two-sided binomial test), and gene expression regulation (3 observed vs. 0.03 expected, >100-fold enrichment, $p_{adj} = 2.42 \times 10^{-2}$, two-sided binomial test). There is also significant enrichment for genes involved in the presynapse (5/336 genes; $p = 3.74 \times 10^{-4}$, OR = 9.25 [3.40-Inf], one-sided Fisher's exact test) and postsynaptic density (11/1,755 genes; $p = 2.17 \times 10^{-4}$, OR = 4.50 [2.26-Inf], one-sided Fisher's exact test), and targets of FMRP (14/842 genes; $p = 1.20 \times 10^{-10}$, OR = 14.37 [7.57-Inf], one-sided Fisher's exact test).

In addition to recurrent mutations within the protein-coding portion of genes, we also assessed codons in which two or more *de novo* missense mutations in unrelated individuals with NDDs have been identified, hereafter referred to as sites. We identified 40 sites in 36 genes, 10 of which have a significant burden of *de novo* missense mutation, after excluding mutations observed in population controls (minor allele frequency (MAF) > 0.01% in the Exome Sequencing Project (ESP; NHLBI GO ESP Exome Variant Server, Seattle, WA (http://evs.gs.washington.edu/EVS/) [August 2016]) (N = 6,503) or the Exome Aggregation Consortium[93] (ExAC) database v.0.3 without neuropsychiatric disorders (N = 45,376)) (**Appendix 2.3**). None of these mutations were observed in unaffected controls in denovo-db v.0.9. Seven sites had more than two recurrent mutations (e.g., PACS1 with six substitutions at residue 203) and some genes had more than one recurrently mutated codon (e.g., *SCN2A*). Sixteen of the sites involved adjacent mutations in the same codon. Twenty-eight of the 40 sites (36/56 mutations) involve CpG dinucleotides, consistent with their known association with hotspots of single-nucleotide variation[94]. Thirty-four sites had average CADD scores of 20 or greater and 17 had a score over 30, indicating that

they are in the top 1% of deleterious mutations in the human genome. This observation stands in contrast to the pattern of *de novo* recurrent missense in controls, where only one of the three sites had a CADD score greater than 20, although the number of events compared is few.

2.5 Discussion

The objective of this research study was twofold: define the features of likely disease-causing *de novo* missense mutations and identify new genes and functional domains relevant to the pathology of NDDs. To increase sample size, we broadly defined NDDs to include not only data from patients with ASD, DD, and ID but also patients with epilepsy and schizophrenia due to the extensive comorbidity of these diagnoses. As expected, both recurrence and severity of missense mutations are critical features. The likelihood of a pathogenic mutation rises significantly when three or more missense mutations are observed in a gene ($p = 1.06 \times 10^{-18}$, two-sided Wilcoxon rank-sum test) and, in particular, when the severity of the missense mutation exceeds a CADD score of 28 (>1.2 positive likelihood ratio). We use these features to identify 35 genes with an excess ($q < 0.05$) of *de novo* missense mutations (**Appendix 2.1**). While many of the top-scoring genes are associated with known syndromic and non-syndromic forms of NDD (e.g., *SCN2A* with ASD[67], *PACS1* with Schuurs-Hoeijmakers syndrome[92], and *ALG13* with epilepsy[72]), seven of these candidates have not been previously reported in ClinVar or OMIM.

Among the 35 genes with a significant excess of recurrent missense mutations, 37% (N = 13) have not yet been associated with a *de novo* LGD mutation (e.g., *COL4A3BP*, *PPP2R5D*) suggesting that LGD events are either not tolerated or associated with a different diagnostic outcome. In support of this observation, 71% (N = 25) of genes were also recently highlighted as likely pathogenic in an exome sequencing study of 3,287 individuals with DD[91]. These independent results support the facility of our burden model in identifying potential risk genes.

Further support for the role of these newly identified genes in NDDs comes from assessment of gene function. Accumulating evidence supports a link between the development and function of excitatory

synapses in NDD and ASD[95]. Consistent with this, we find a 35-fold enrichment of genes regulating postsynaptic membrane potential in genes that carry a significant burden of *de novo* missense events. However, with the number of mutations in denovo-db, overall gene burden is insufficient in identifying the complete set of genes that contribute to NDDs via missense mutation. As 13% of missense events seen in individuals with NDDs are causative, we expect that 755 of the 5,807 mutations in denovo-db are implicated. Yet only 218 events occur in the set of genes. While the whole-gene burden model may be sufficient when studying LGD mutations, which are expected to have equivalent truncating effects, additional methods are needed to distinguish pathogenic missense mutations from incidental ones.

The importance of specific missense mutations in disease is evidenced by Schuurs-Hoeijmakers syndrome[92], in which an arginine to tryptophan substitution at residue 203 of *PACS1* is responsible for a well-defined neurodevelopmental phenotype. No other pathogenic SNVs have been described in this gene to date. In addition to this known pathogenic event, 39 other amino acids are substituted in multiple unrelated individuals in denovo-db. Of the 36 genes with a recurrent site, only 10 have a significant burden of events. To better assess the association of these events with disease, we modified our burden model to predict the number of events expected at individual codons. This increased granularity better represents the variability in functional importance of residues across a protein and the resulting variability in missense mutation deleteriousness. Using this model, seven sites have a burden of missense mutation. Six of these genes are also significant when assessing whole-gene burden, including PACS1. However, a known pathogenic event in *ALG13* that was identified with the codon-specific model was missed by the whole-gene model. Given the prior of *de novo* mutations in cases and absence of events in controls, we expect that more of these sites are associated with NDDs. Further investigation into the functional effect of these substitutions will provide insights into the proteins' roles in normal and abnormal development.

Chapter 3: Targeted sequencing of *de novo* missense mutation hotspots reveals novel risk genes.

This chapter has been adapted from: Geisheker, M.R., Heymann, G., Wang, T., Coe, B.P., Turner, T.N., Stessman, H.A.F., et al. (2017). Hotspots of missense mutation identify neurodevelopmental disorder genes and functional domains. *Nature Neuroscience*, 20(8):1043-1051. doi:10.1038/nn.4589.

3.1 Summary

De novo missense mutations in individuals with NDDs occur with both an increased rate and predicted deleteriousness than events in controls, yet few genes reach statistical significance. We focused on recurrent sites and clustering, both seen in NDD cases, as potential patterns of mutation that are associated with disease. To further test this association, we performed targeted sequencing with molecular inversion probes in 17,688 patients with idiopathic NDDs. We targeted 20 recurrent sites initially, and later also targeted regions with multiple mutations in close proximity. We identified 21 new patients with missense mutations identical to those seen in our discovery set. One recurrent substitution (p.Ala636Thr) occurs in a glutamate receptor subunit, GRIA1. This same amino acid substitution corresponds precisely to a mutation in the homologous but distinct mouse glutamate receptor subunit Grid2, associated with Lurcher ataxia. Phenotypic follow-up in five individuals with *GRIA1* mutations shows evidence of specific learning disabilities and autism, and functional studies confirm a dominant negative effect of this mutation in the third transmembrane domain of the receptor. We also observed variants of interest nearby to site mutations, and find significant clustering of *de novo* missense mutations in 200 genes. These results highlight specific functional domains and synaptic candidate genes important in NDD pathology.

3.2 Introduction

Comparison of *de novo* missense mutations between cases and controls in denovo-db (v.0.9) provides further evidence, both quantitative and qualitative, of the contribution of *de novo* missense mutations in neurodevelopmental disorders (NDDs). Burden analysis identified 35 genes with more *de novo* missense mutations than expected, but the number of mutations in these genes is less than a third of the number of events that are expected to be pathogenic[28]. Due to the variable nature of missense mutations, we hypothesize that specific missense mutations are pathogenic in NDDs. Assessment of overall gene burden, which characterizes each event as equitable in effect, overlooks many of these events and a more focused approach is likely to identify novel deleterious mutations.

The disease risk of some *de novo* substitutions, such as p.Arg203Trp in PACS1, is known. Yet only seven events in denovo-db[79] (v.0.9) reach statistical significance. This is likely caused by the large locus heterogeneity in NDDs, first predicted with early work on CNVs and affirmed with continuing studies of SNVs[17,18,25,47,96]. With missense mutations, the high incidental rate adds further challenge to the identification of these specific events[28]. These challenges can be surmounted by studying a larger number of individuals with NDDs. While the costs associated with exome sequencing may be prohibitive, targeted sequencing is an affordable, efficient way to survey regions of interest[46,97,98]. It will enable us to test the association of specific missense mutations with NDDs, and guide functional testing towards events that have molecular and clinical evidence of potential deleteriousness.

3.3 Methods

Targeted sequencing. Single-molecule molecular inversion probes (smMIPs[43]) were designed with the MIPgen program[97] to capture sequences of interest. To maximize coverage, we designed one smMIP for each strand for each target. We first used smMIPs targeting 24 sites to sequence eight cohorts containing a total of 6,058 cases and 2,854 controls (**Appendices 3.1** and **3.2**). We also used smMIPs targeting two codons thought to be sites but later discovered to be duplicate database entries (*TBR1*) or present in both the case and her unaffected sister (*PDCD11*). As clusters of missense mutations have been associated

with NDDs[99], we then designed a set of smMIPs targeting 17 clusters in denovo-db v.0.9. These cluster smMIPs, along with the 24 site smMIPs, were used on an additional four cohorts, containing 5,055 cases and 169 controls. A final set of smMIPs was created that excluded those targeting four sites that had no brain expression (*AGER*, *ZNF215*), low CADD scores (*ALDH5A1*), or high frequency in control populations (*DUSP15*). This final set, targeting 20 sites and 17 clusters, was used on five new cohorts, containing a total of 6,576 cases. Across all three designs, this totals to 17,688 cases and 3,023 controls (**Appendices 3.1** and **3.2**). The study size was not predetermined but based on the maximal number of samples that could be screened. Reads were aligned using BWA-MEM[100] to GRCh37/hg19. All 146 rare (MAF < 0.01%) variants with CADD score >20 were validated with Sanger sequencing (**Appendix 3.3**). Patients were initially identified through targeted sequencing in anonymized ASD and DD cohorts. All patients were consented for sequencing and recontacting for inheritance testing at the providing laboratory. Patient samples were acquired from Adelaide (Jozef Gecz, University of Adelaide), Antwerp (Frank Kooy, University of Antwerp), Autism Clinical and Genetic Resources in China (ACGC; Kun Xia), the Autism Genetic Resource Exchange (AGRE), Iowa (Jacob Michaelson, University of Iowa), Leiden (Gijs Santen, Leiden University Medical Center), Leuven (Hilde Peeters, University Hospitals Leuven), Philadelphia (Hakon Hakonarson, Children's Hospital of Philadelphia), Prague (Zdeněk Sedláček, University Hospital Motol), San Diego (Eric Courchesne, UC San Diego), Simons Simplex Collection (SSC), Stockholm (Magnus Nordenskjold, Karolinska University Hospital), the Study of Autism Genetics Exploration (SAGE; Raphe Bernier, University of Washington), The Autism Simplex Collection (TASC), and Troina (Corrado Romano, Associazione Oasi Maria Santissima).

***GRIA1* transfection and patch-clamp recording assays.** *GRIA1* wild-type and A636T mutant DNA sequences were synthesized (GenScript) and cloned into mammalian expression vectors. Human embryonic kidney (HEK) 293T/17 SF (ATCC ACS-4500) cells, routinely used for transient transfection and electrophysiological recordings as they allow robust heterologous expression, were cultured in DMEM (Invitrogen) supplemented with 10% Fetal Bovine Serum and 1% Streptomycin up to a maximum passage number of 15. For transient transfection, cells were split and plated onto 12 mm glass coverslips (Carolina Scientific) coated with Poly-L-Lysine (50 ng/µl). Then, 4-6 hours later, approximately 0.6 µg of

total DNA/coverslip was transfected using the Fugene6 reagent (2 µl/coverslip, Promega). For heteromeric cells, approximately 0.3 µg of WT and A636T were co-transfected. Whole-cell recordings were performed approximately 60 hours after transfection using a Multiclamp 700B amplifier (Molecular Devices) with glass micropipettes of resistance 2-5 MΩ. Extracellular solution contained (in mM): 150 NaCl, 2.5 $CaCl_2$, 2.5 KCl, 1 $MgCl_2$, 10 D-Glucose, 10 HEPES, pH to 7.4 with NaOH. Intracellular pipette solution contained (in mM): 140 CsCl, 2 $MgCl_2$, 10 HEPES, 10 EGTA, pH to 7.3 with CsOH. Voltage-ramp recordings ranged from -100mV to 80 mV and spanned 1.8 seconds. Data were collected with sampling at 10 kHz and only cells with whole-cell access resistance that remained less than 15 MΩ across recordings were included in analysis. To verify channel expression, a saturating concentration of glutamate (1 mM) was applied with 100 µM CX614, and only cells with detectable current were included. NBQX and CX614 were acquired from Tocris Biosciences. Sample size was chosen based on previous literature and variance of ion channel studies of similar nature.

Array comparative genomic hybridization. Array labeling and hybridization was performed as previously described[101]. Briefly, 250 ng of sample DNA was labeled with Cy3 using a NimbleGen labeling kit (Roche). Reference DNA (NA12878) was labeled in a pooled reaction for four arrays with Cy5 using 1 ug of DNA. Hybridization was performed using the Agilent 2x400K array platform using standard reagents, imaged using an Agilent Scanner, and processed using Agilent Feature Extraction. CNV calls were generated using Agilent CytoGenomics 4.03.12 and the ADM2 calling algorithm with default parameters. For samples passing standard Agilent QC parameters (DLRSD < 0.2), all CNVs over 100 kbp were visually inspected, filtered for known reference sample artifacts, and compared to 29,085 cases of ID/DD and 19,584 controls[102] to identify rare CNVs that may contribute to pathogenicity in these cases.

Missense clustering. Genes with significant clustering of missense mutations were identified by CLUMP[51] (CLUstering by Mutation Position; https://github.com/karchinlab/clump), which applies an unsupervised clustering algorithm based on partitioning around medoid distances between mutations. We implemented the permutation (-z 1000) and minimum mutation options (-m 2) and calculated a p-value based on the null distribution of case and control CLUMP score differences. The case set included

individuals with an NDD primary phenotype (ASD, DD, ID, or EPI) from denovo-db[79] v.1.2 (**Appendix 3.4**) and consisted of 22 studies[28,39,44,56,57,59,61–63,65–70,72–76,85,103] with 9,997 affected individuals (8,917 *de novo* missense variants). We compared against two control missense datasets: 1) missense mutations (MAF < 1%) from Europeans[51] (N = 420; 196,260 mutations) from the 1000 Genomes Project[104] and 2) private missense mutations present in individuals from the ExAC v.0.3 without neuropsychiatric disorders (N = 45,376; 1,466,439 mutations)[93]. All variants were re-annotated using the CRAVAT software to enable exact transcript comparisons[105].

3.4 Results

Targeted sequencing of missense mutations. Using single-molecule molecular inversion probes (smMIPs), we targeted 20 of these recurrent sites for sequencing in a large cohort of 17,688 patients with a primary diagnosis of ASD or DD/ID (**Appendices 3.1** and **3.2**). The set included primarily patients with idiopathic NDDs not yet tested by exome sequencing. We also included a set of unaffected siblings as an additional control (N = 3,023). We identified and validated 21 recurrent missense variants at 12 sites in 11 genes among cases (**Table 3.1**, **Figure 3.1a-c**). No variants were observed at any of the 20 sites in controls. The inheritance status for only eight of the events identified in cases could be determined due to missing parental DNA—six were determined to be *de novo* (**Table 3.1**; PACS1 p.Arg203 (two substitutions), GRIA1 p.Ala636, SCN2A p.Arg937, and SMAD4 p.Ile500 (two substitutions)). Interestingly, one of the inherited substitutions (PTPN11 p.Gly503Glu) is adjacent to the well-known Noonan syndrome recurrent substitution[106] (PTPN11 p.Ser502Thr) and was transmitted paternally to two children both affected with ASD and ID. No information on the father's phenotype is currently available. Five genes corresponding to six sites were identified with two or more recurrent missense mutations in the NDD cohort, namely GRIA1 p.Ala636, PACS1 p.Arg203, SCN2A p.Arg379, SCN2A p.Arg937, SMAD4 p.Ile500, and ZNF215 p.Arg473. Phenotypic similarities are present in patients with shared mutations, such as ALG13 (**Figure 3.1b**), where all six individuals with a mutation at residue 107 have both EPI and DD even though they were recruited from cohorts with different primary diagnostic criteria. Both

individuals with newly found mutations at SMAD4 p.Ile500 have features consistent with Myhre syndrome, including ID, short stature, facial dysmorphisms, and hearing loss[107,108] (**Appendix 3.5**).

Table 3.1 New recurrent mutations at targeted missense sites.

Gene	Site	Alternate amino acid(s)	Protein ID	Mutations in denovo-db v.0.9 (N = 8477)	Codon *de novo* p	Codon *de novo* p, genome-wide correction*	Mutations identified with smMIPs (N = 17850)			Total *de novo*	Codon *de novo* p	Codon *de novo* p, genome-wide correction*	ExAC v.0.3 allele count (N = 45376)
							Inherited	*De novo*	Unknown				
PACS1	p.Arg203	Trp	NP_060496.2	6	1.03E-24	1.13E-17		2 (St, Tr)	1 (Ad)	8	1.37E-29	1.51E-22	1
PPP2R5D	p.Glu198	Lys	NP_006236.1	4	3.84E-18	4.22E-11			1 (Tr)	4	3.48E-16	3.83E-09	0
ALG13	p.Asn107	Ser, Thr	NP_060936.1	5	5.26E-17	5.79E-10			1 (Tr)	5	1.47E-14	1.62E-07	0
SCN2A	p.Arg937	Cys, His	NP_001035232.1	3	9.47E-12	1.04E-04		1 (ACGC)	1 (ACGC)	4	8.26E-14	9.09E-07	0
SMAD4	p.Ile500	Val, Thr	NP_005350.1	2	1.12E-07	1		2 (An, Le)		4	1.88E-13	2.07E-06	0
PTPN11	p.Gly503	Arg, Glu	NP_002825.3	3	1.66E-11	1.83E-04	1† (AGRE)			3	5.10E-10	5.62E-03	0
GRIA1	p.Ala636	Thr	NP_000818.2	2	1.11E-07	1		1 (St)	2 (St)	3	4.89E-10	5.39E-03	0
SCN2A	p.Arg379	His	NP_001035232.1	2	7.39E-08	8.14E-01			1 (Ad)	2	7.04E-07	1	0
CLCN4	p.Arg718	Trp, Gln	NP_001821.2	2	9.57E-08	1	1 (St)			2	9.11E-07	1	0/2††
KCNQ3	p.Arg230	Cys	NP_004510.1	2	3.36E-07	1			1 (Tr)	2	3.20E-06	1	0
ZNF215	p.Arg473	Gln	NP_037382.2	2	6.54E-07	1			3 (Ad)	2	3.49E-06	1	5
CUX2	p.Glu590	Lys	NP_056082.2	2	5.38E-07	1			1 (Tr)	2	5.12E-06	1	0

*Bonferroni family-wise error rate (FWER) correction based on 1.1E7 codons in genome.

†In two affected siblings.

††Allele in denovo-db v.0.9 has 0 occurrences in ExAC; allele found with smMIPs has been seen twice.

ACGC, Autism Clinical and Genetic Resources in China; Ad, Adelaide; AGRE, Autism Genetic Resource Exchange; An, Antwerp; Le, Leuven; St, Stockholm; Tr, Troina

We also observed rare, potentially disruptive, missense variants in close proximity to the original recurrent site mutations, such as in *SMAD4* (**Figure 3.1c**). We reexamined our database for regions where multiple *de novo* substitutions mapped within 10 amino acids. We designed smMIPs for 17 clustered regions as well as the 20 recurrent sites (in 30 total genes) and sequenced this extended set (~5 kbp of coding sequence) in a subset of the NDD cohort (**Appendix 3.1**). Combined with targeted sites, we discovered a total of 139 recurrent or clustered missense variants in 137 cases compared to seven variants in five unaffected siblings, representing a significant enrichment ($p = 1.11 \times 10^{-4}$, OR = 3.93 [1.76-10.89], two-sided Fisher's exact test) (**Table 3.2** and **Appendix 3.3**). Twelve of the clustered missense mutations in cases were confirmed *de novo*, including events in *SATB2* (**Figure 3.1d**), *GRIA1* (**Figure 4a**), *SCN2A*, *KCNQ3*, *SCN8A*, *DEAF1*, and *PPR2R1A* (**Appendix 3.6**).

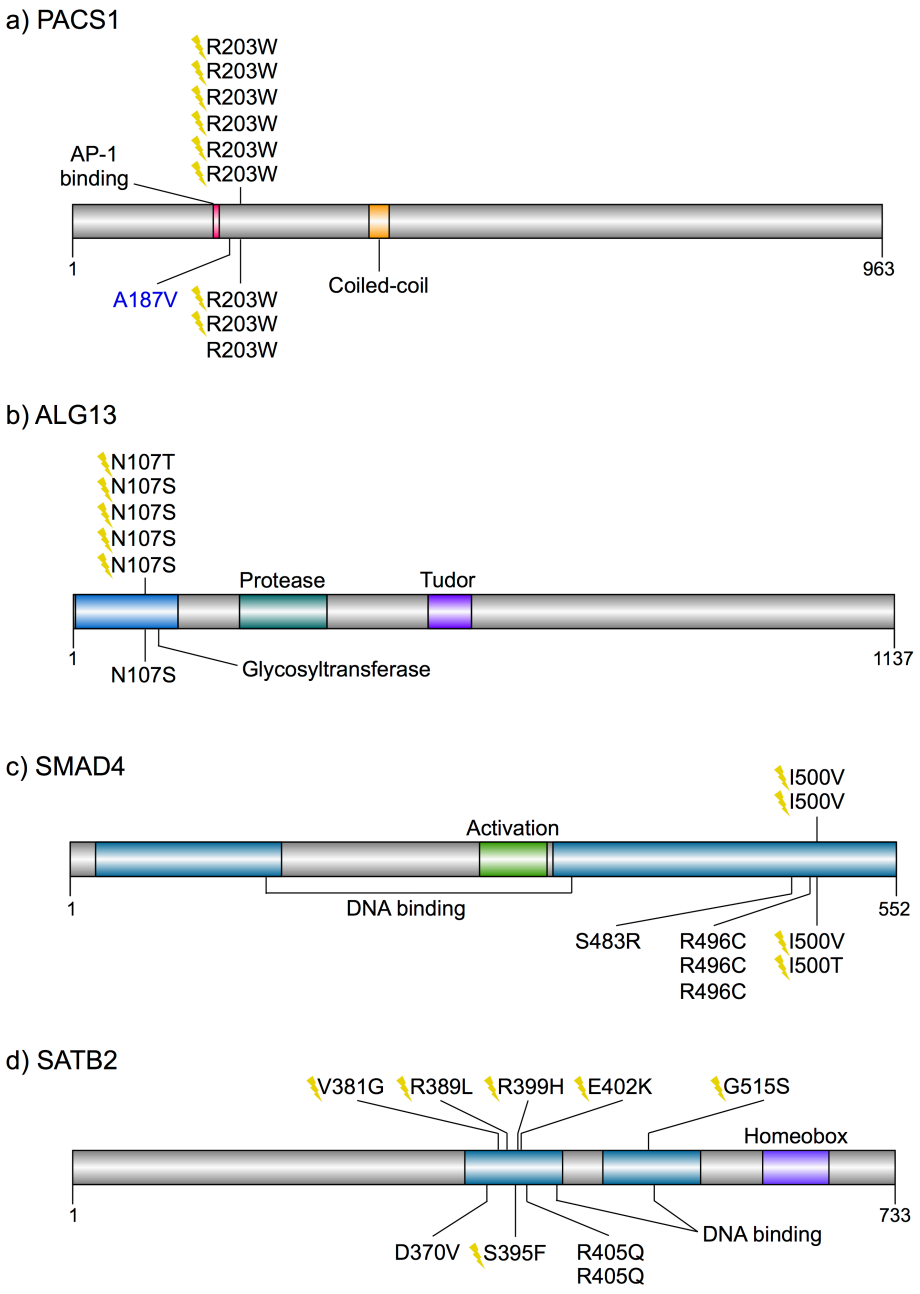

Figure 3.1. Recurrent substitutions fall in or near functional domains. Published substitutions in NDD patients are above the protein and new substitutions identified by targeted sequencing are below the protein. *De novo* events (lightning bolt) and paternally inherited events (blue) are indicated. Inheritance is unknown for the remaining events. Protein domains are from UniProt. **a)** PACS1, NP_060496.2. **b)** ALG13, NP_060936.1. **c)** SMAD4, NP_005350.1. **d)** SATB2, NP_001165988.1.

Table 3.2. Rare (MAF < 0.01%) clustered missense mutations identified by targeted sequencing.

				Controls		Cases		All cases**		
Gene	Site or cluster	Protein ID	denovo-db v.0.9	N	Missense	N	Missense	N	Missense	Known *de novo*
PACS1	p.Arg203	NP_060496.2	6	3023	0	17689	4	26166	10	8
PPP2R5D	p.Glu198	NP_006236.1	4	3023	1	17689	3	26166	7	4
SCN2A	p.Arg937	NP_001035232.1	3	3023	0	17689	5	26166	8	5
DEAF1	p.Gln264	NP_066288.2	2	3023	0	17689	7	26166	9	5
ALG13	p.Asn107	NP_060936.1	5	3023	0	17689	1	26166	6	5
GRIA1	p.Ala636	NP_000818.2	2	3023	0	17689	5	26166	7	4
COL4A3BP	p.Ser260	NP_001123577.1	3	3023	0	17689	2	26166	5	3
SCN8A	p.Gly214-p.Asn215	NP_055006.1	3	169	0	11631	1	20108	4	3
DEAF1	p.Leu219-p.Gly220	NP_066288.2	2	169	0	11631	2	20108	4	3
SATB2	p.Arg399-p.Glu402	NP_001165988.1	2	169	0	11631	3	20108	5	3
PPP2R1A	p.Arg182	NP_055040.2	2	3023	0	17689	2	26166	4	3
SCN8A	p.Arg1617-p.Gly1625	NP_055006.1	2	169	0	11631	2	20108	4	3
PTPN11	p.Gly503	NP_002825.3	3	3023	0	17689	6	26166	9	3
STXBP1	p.Gly544	NP_001027392.1	2	3023	0	17689	5	26166	7	3
BCL11A	p.Thr47-p.Cys48	NP_075044.2	2	169	0	11631	4	20108	6	2
KCNQ3	p.Arg230	NP_004510.1	2	3023	0	17689	3	26166	5	3
TRRAP	p.Trp1848	NP_003487.1	2	3023	2	17689	9	26166	11	3
CLTCL1	p.Val641-p.Asn647	NP_009029.3	2	169	0	11631	2	20108	4	2
TRPM7	p.Thr379-p.Glu387	NP_060142.3	2	169	0	11631	2	20108	4	2
SCN2A	p.Arg853	NP_001035232.1	2	3023	1	17689	2	26166	4	2
DUSP15	p.Thr4	NP_001012662.1	2	3023	0	11113	3	19590	5	2
SCN2A	p.Arg379	NP_001035232.1	2	3023	0	17689	6	26166	8	2
SATB2	p.Val381-p.Arg389	NP_001165988.1	2	169	0	11631	1	20108	3	2
SMAD4	p.Ile500	NP_005350.1	2	3023	0	17689	6	26166	8	4
SLC35C2	p.Ser173-p.Gly176	NP_057029.8	2	169	0	11631	3	20108	5	2
CLCN4	p.Arg718	NP_001821.2	2	3023	1	17689	4	26166	6	2
PCGF2	p.Pro65	NP_009075.1	2	3023	0	17689	5	26166	7	2
KAT6B	p.Ser1380-p.Glu1389	NP_001243398.1	2	169	0	11631	4	20108	6	2
TRIO	p.Pro1461	NP_009049.2	2	3023	0	17689	1	26166	3	2
NCAN	p.Pro1219-p.Val1221	NP_004377.2	2	169	0	11631	8	20108	10	2
ITPR1	p.Thr267-p.Arg269	NP_001161744.1	2	169	0	11631	1	20108	3	2
CUX2	p.Glu590	NP_056082.2	2	3023	1	17689	3	26166	5	2
ZNF215	p.Arg473	NP_037382.2	2	3023	1	11113	10	19590	12	2
SMARCA2	p.Arg525	NP_003061.3	2	3023	0	17689	3	26166	5	2
TBR1	p.Trp271	NP_006584.1	1	3023	0	17689	3	26166	4	1
PDCD11	p.Arg964	NP_055791.1	1	3023	0	17689	8	26166	9	1
		TOTAL	**83**		**7**		**139**		**222**	**101**

*Minor allele frequency (MAF) < 0.01% in ExAC v.0.3 and ESP v.0.0.30.

**denovo-db v.0.9 and smMIPs.

In addition to new variants at sites in denovo-db v.0.9, targeted sequencing established 14 new sites, although inheritance status for most variants remains unknown. The specific substitutions at SCN8A p.Arg1617 and STXBP1 p.Arg551 have been seen previously in NDD cases. While Myhre syndrome has been associated only with residue 500 of SMAD4[108], *in silico* predictions suggest that the three p.Arg496Cys substitutions we identified are also likely to be pathogenic as the residue is highly conserved across species and the amino acid substitution is nonconservative[109]. Detailed phenotypic

information on one patient with this mutation indicates characteristics of the syndrome, including ID, short stature, and dysmorphic facial features, suggesting that Myhre syndrome is not only limited to one codon[108]. Phenotypic commonalities are also present amongst individuals with clustered mutations, indicating the functional relevance of protein domains. For example, seven out of eight patients with a substitutions in the first DNA binding domain of SATB2 (**Figure 3.1d**) have facial dysmorphisms and seven out of eight have DD.

De novo* missense mutations in *GRIA1. We identified a recurrently mutated codon in *GRIA1* (protein product GluA1; **Figure 3.2a**), a subunit of AMPA glutamate receptors, which was originally reported in one patient with ID[74] and another with ASD[75]. Both patients share an identical *de novo* G>A mutation resulting in an alanine to threonine amino acid replacement at residue 636 (NP_000818.2). Resequencing identified the same variant in three more patients with a primary diagnosis of ASD. One newly found mutation was confirmed as *de novo*; paternal DNA is not available for the other two but the mutation is not present in either of the patients' mothers. Using array comparative genomic hybridization, we found no evidence for large pathogenic CNVs in any of the three patients for whom we had DNA. While this position is a CpG dinucleotide and therefore prone to recurrent mutation, this variant has not been observed in 60,706 individuals published by ExAC[93]. Moreover, we identified a second *de novo* missense mutation in close proximity (**Figure 3.2a**) in a patient with DD. The dearth of variants in healthy controls and the observation of the same recurrent variant in six unrelated patients (three of which were *de novo* ($p = 5.39 \times 10^{-3}$, one-tailed binomial test, genome-wide correction)) suggested that the mutation was pathogenic.

The mutated site maps to the eighth position (p.Ala636) of a highly conserved 9-amino acid motif, SYTANLA**A**F (**Figure 3.2b**), present in the M3 transmembrane domain of all glutamate receptors, which plays a critical role in channel gating[110]. The specific alanine to threonine substitution observed in the five patients here has been observed at the functionally equivalent site in other members of the glutamate receptor gene family. It was first identified as a spontaneous mutation in *Grid2* in a mouse line at Jackson Laboratories (Lurcher) that results in a constitutively active channel comprised of homomeric GluRδ2

subunits selectively expressed in cerebellar Purkinje neurons[111]. Mice heterozygous for this substitution in the GluRδ2 receptor develop severe ataxia as a consequence of neurotoxicity from excess current flux. Notably, humans with the substitution in GluRδ2 also suffer from ataxia[112]. Engineering of the A>T mutation at the homologous site in the rat isoform of the GluA1 receptor produces a similar constitutively active phenotype with altered kinetic and pharmacological properties[113–115].

To confirm constitutive activity or leak current in the human isoform of *GRIA1* identified in affected patients, we synthesized cDNA encoding the human wild-type (WT) and mutant (A636T) at base-pair position 1906 (G/A). Leak current was measured using whole-cell voltage-clamp recordings of HEK 293 cells heterologously expressing either WT or A636T in the absence of agonist by applying a voltage ramp from -100mV to +80mV. GluA1-mediated current was confirmed by application of the AMPA receptor-selective antagonist 2,3-dihydroxy-6-nitro-7-sulfamoyl-benzo[f]quinoxaline-2,3-dione (NBQX), followed by an additional voltage ramp. Subtracted current in the presence of NBQX revealed a notable constitutive current in A636T, but not WT-expressing cells (**Figure 3.2d,f**). Consistent with GluA1-mediated current, inward rectification is abolished following channel blockade with NBQX. No changes in current magnitude or shape were seen in cells expressing the WT channel after NBQX application (**Figure 3.2c,f**). Affected patients with the p.Ala636Thr substitution are heterozygous indicating that a majority of receptors are likely comprised of WT and p.Ala636Thr receptor subunits. To assess the functional phenotype of these 'heteromeric' receptors, we co-transfected equal ratios of WT and A636T DNA and performed the same voltage-ramp recordings (**Figure 3.2e**). While a noticeable constitutive current was still present, it was smaller than the homomeric p.Ala636Thr channel demonstrating that the overall effects of the mutation are mitigated by the presence of the WT subunits (**Figure 3.2f**).

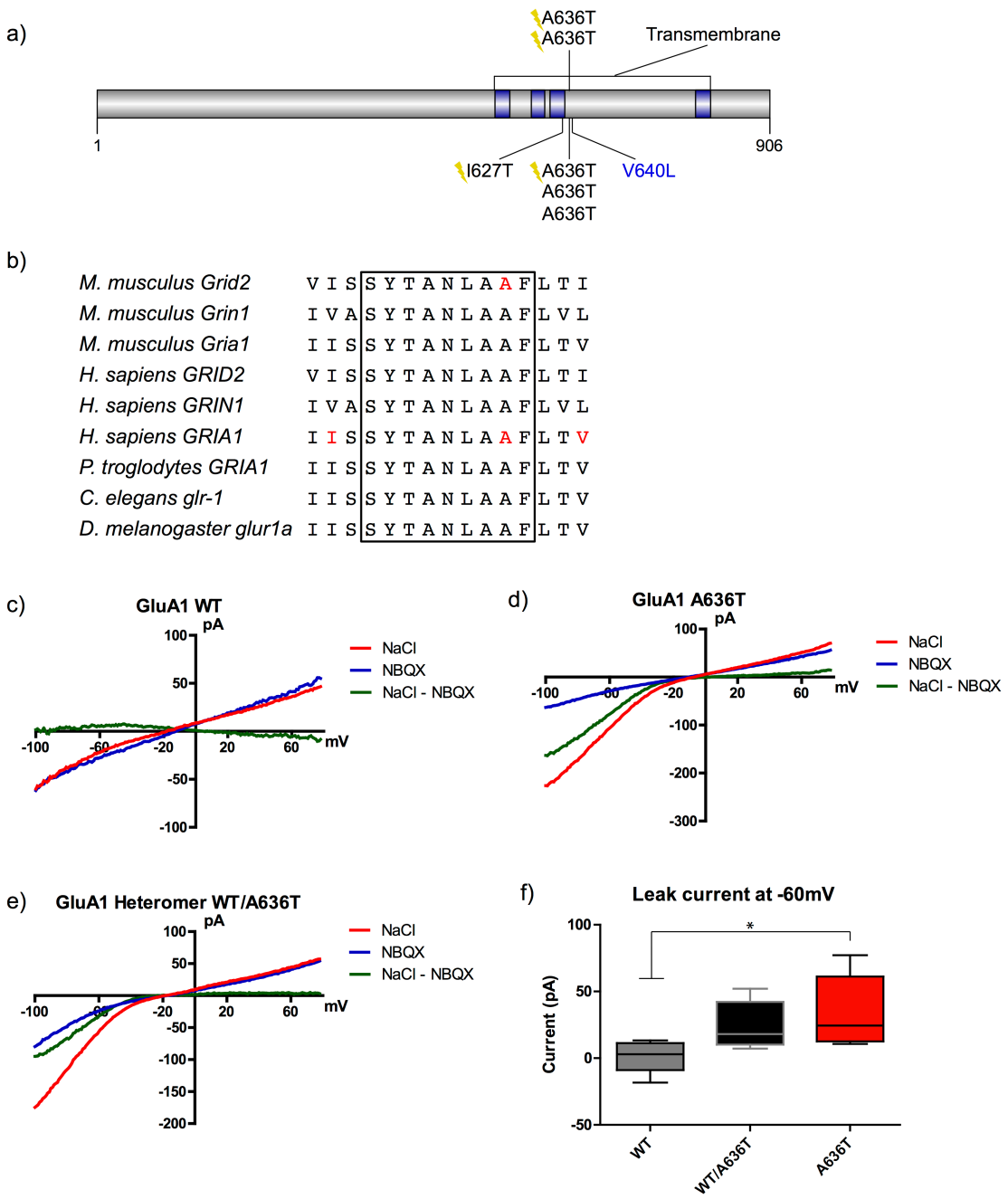

Figure 3.2. Functional effect of recurrent GRIA1 missense mutations. a) Linear representation of annotated domains in the protein GRIA1, a.k.a. GluA1 (NP_001244950.1) as defined in UniProt. **b)** A recurrent substitution observed only in NDD cases (N = 5 patients) falls within a highly conserved M3 transmembrane domain, important in channel gating. The alanine substituted in these patients (red) is homologous to the one that causes severe ataxia in Lurcher mice in the delta-2 subunit of this receptor (GRID2). **c-e)** Example current traces from a 1.8s voltage-ramp from -100mV to +80mV for **c)** WT, **d)** A636T, and **e)** heteromeric WT/A636T transfected HEK cells. The three current traces per panel correspond to voltage-ramp currents in the presence of normal extracellular solution (NaCl), extracellular solution supplemented with 50 uM NBQX (NBQX), and the isolated GluA1 dependent current determined by subtracting the NBQX current from the NaCl current (NaCl – NBQX). **f)** Average leak current at -60mV. The GluA1-mediated current (NaCl – NBQX) was determined at -60mV and averaged across cells (N = 5 (WT), 7 (A636T), and 5 (heteromeric); p = 0.024, F = 4.91, 2 degrees of freedom, one-way ANOVA). Data are mean +/- S.E.M.

Consistent with the prevalent role of GluA1 homomeric channels in synapse development and synaptic plasticity[116], phenotypic analysis of four of the individuals with the p.Ala636Thr substitution demonstrates common features (**Appendix 3.7**), including mild to moderate ID (4 of 4 individuals) and ASD (3 of 4 individuals). Three of the four for whom information is available had delayed language development, with two (both with ASD) demonstrating persistent difficulties with pronunciation and vocabulary. These two individuals were also noted to have highly similar facial features and were diagnosed with ADHD. Similarly, the individual without ASD is noted to have behavioral dysfunction. Two individuals also had delayed motor development. All four have normal MRIs. Collectively our evidence suggests that this specific missense mutation dictates a common pathological brain development trajectory and supports the idea that specific substitutions contribute to NDD pathogenesis.

Clustered missense mutations and functional domains. Our sequencing results as well as the *GRIA1* analysis strongly suggest that clustered and recurrent missense mutations have the potential to highlight functional protein domains important in NDD pathology. We previously developed a tool, CLUMP[51], to assess the significance of clustered mutations and we applied it to an updated version of denovo-db (v.1.2) to identify genes and functional domains for future investigation. Overall, we examined 8,917 *de novo* missense mutations in cases and calculated raw CLUMP scores for 1,699 proteins containing at least two mutations in cases. We performed case–control analyses comparing the pattern of private alleles in ExAC and separately among European individuals from the 1000 Genomes Project (1KG). Twenty-eight out of 34 genes we initially identified were testable by this approach and 18 of them showed nominally significant clustering of *de novo* missense mutations ($p < 0.05$, CLUMP, one-tailed permutation test). Altogether, we identified 200 genes with significant clustering of missense mutations at the protein level (**Appendix 3.8**). Once again, this set is significantly associated with aspects of neuronal communication, including regulation of the postsynaptic potential (11 observed vs. 1 expected, 11.0-fold enrichment, $p_{adj} = 6.93 \times 10^{-5}$, two-sided binomial test) and synaptic signaling (20 observed vs. 4.15 expected, 4.82-fold enrichment, $p_{adj} = 8.38 \times 10^{-5}$, two-sided binomial test), as well as chromatin-mediated maintenance of transcription (4 observed vs. 0.1 expected, 40.7-fold enrichment, $p_{adj} = 2.96 \times 10^{-2}$, two-sided binomial test). Many of the genes encode channel proteins and receptors (e.g., *GRIA1*, *GRIN1,*

GRIN2A, *GRIN2B*, *KCNH1*, *KCNQ2*) and exhibit clustering in or near specific functional domains, such as the transmembrane, pore or voltage sensor domains (**Figure 3.3a-d**). Other proteins, such as CTCF, are remarkable in that the clustering pattern of substitutions in patients highlights a subset of the C2H2 ZNF motifs, which are never substituted in controls (**Figure 3.3e**). These pockets of patient-only missense mutations will be increasingly important in characterizing pathogenic genes and functional domains.

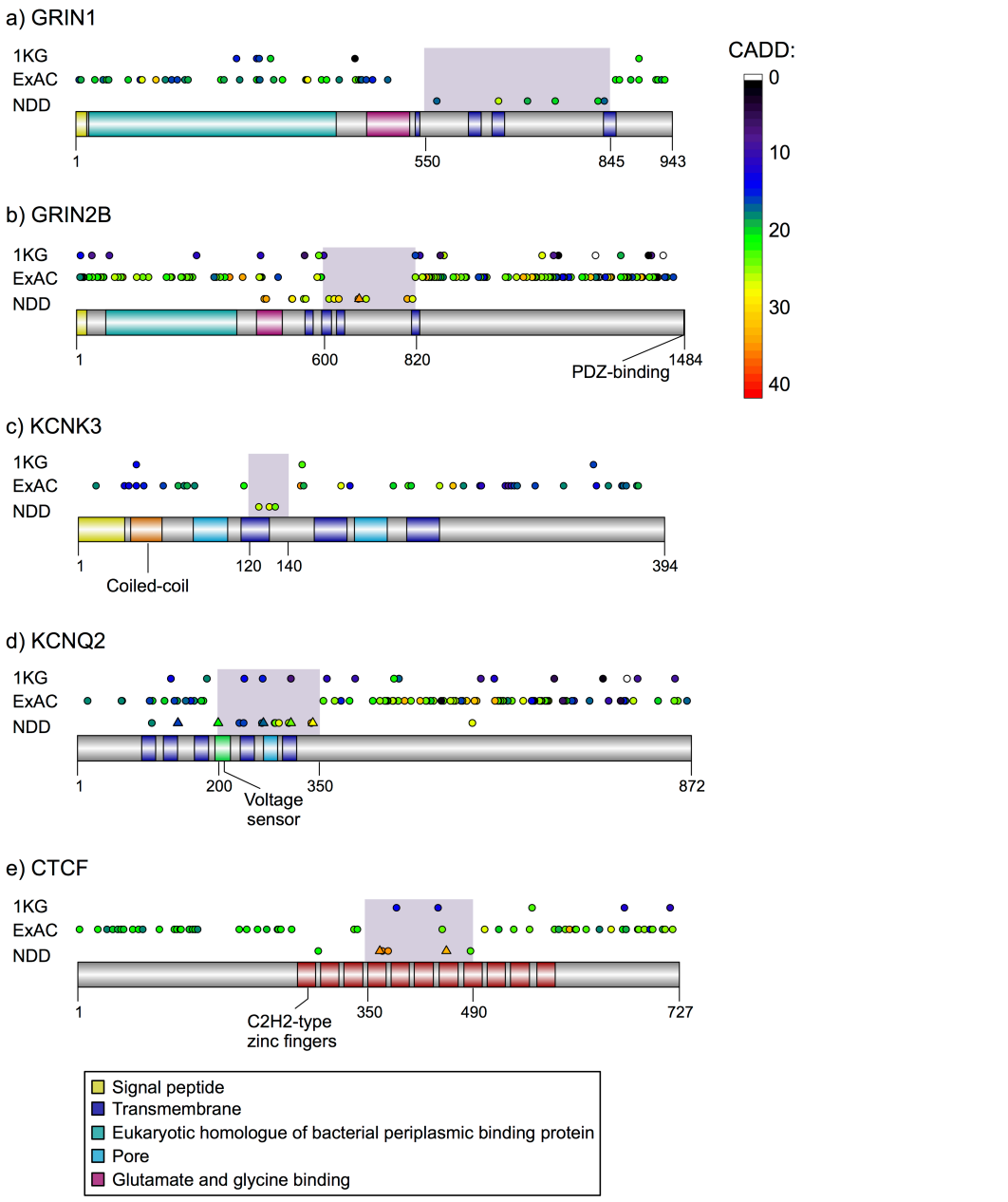

Figure 3.3. Proteins with excessive clustering of missense mutations in NDD cases. The pattern of *de novo* missense mutations in cases with NDDs is contrasted with rare missense variants from 1KG and private missense mutations from ExAC excluding neuropsychiatric cases. Substitutions are colored by severity (CADD heatmap) and recurrent site substitutions are indicated (triangle). Significance of clustering was calculated based on comparison to ExAC using CLUMP. **a)** GRIN1 (NP_001172019.1) shows greater clustering of substitutions in NDD patients (CLUMP = 1.68, p = 0.013) with region-specific significance corresponding to the transmembrane domains (amino acids 550-845; Fisher's exact test $p = 5.6 \times 10^{-8}$). **b)** Similarly, substitutions cluster for GRIN2B (NP_000825.2; CLUMP = 1.34, p = 0.003) in particular between the second and fourth transmembrane domains (amino acids 600-820; $p = 2.0 \times 10^{-9}$). **c)** KCNK3 (NP_002237.1) patient substitutions cluster (CLUMP = 0.54, p = 0.036) near the first transmembrane domain (amino acids 120-140, $p = 9.4 \times 10^{-5}$, OR = Inf). The average per-base rate of ExAC samples with ≥10X coverage across the exon harboring mutations in cases was 79.1%. **d)** KCNQ2 (NP_742105.1) shows several hotspots (CLUMP = 0.36, $p < 1 \times 10^{-3}$) corresponding to the pore and voltage sensor of the channel (amino acids 200-350, $p = 2.0 \times 10^{-14}$). **e)** Finally, patients show more severe CTCF (NP_006556.1) substitutions that cluster (CLUMP = 1.0, p = 0.007) at two locations between the fourth and seventh C2H2 zinc finger motifs (amino acids 350-490, $p = 9.1 \times 10^{-8}$).

3.5 Discussion

Our discovery dataset, denovo-db, shows evidence of the contribution of missense mutations to NDDs, including recurrent site mutations. However, the locus heterogeneity of the disorders requires a larger sample size to confidently associate specific mutations with disease risk. Targeted sequencing of specific protein-coding regions efficiently captured sites of interest and discovered additional variants of interest in close proximity. It showed that recurrent and clustered amino acid replacements are more common in cases than controls ($p = 1.11 \times 10^{-4}$, OR = 3.93 [1.76-10.89], two-sided Fisher's exact test). In addition to finding additional site mutations, new sites of interest were discovered with targeted sequencing, including a missense mutation in *SMAD4*, observed in three cases, that is predicted to be pathogenic.

We also identify 200 genes with patterns of *de novo* missense mutations that are more clustered in cases when compared to population controls (**Appendix 3.8**), 79% (N = 157) of which have not yet been associated with an NDD in OMIM or ClinVar databases. Of the 200 genes with significant clustering of missense mutations, 67% (N = 134) did not show any evidence of LGD mutation in NDDs in denovo-db v.1.2, OMIM, or ClinVar; 45% (N = 89) have been shown to be loss-of-function intolerant in the ExAC database[93], suggesting that LGD mutations in them may be genetically lethal (e.g., cause embryonic lethality or infertility), although additional experiments will be required to make this determination. In many cases, the clustering of *de novo* substitutions highlights protein functional domains (**Figure 3.1**), such as specific zinc-finger motifs (e.g., *CTCF*), transmembrane domains (e.g., *GRIN1*), and voltage sensors and channel pores (e.g., *KCNQ2*). As the number of exomes increases, these hotspots of pathogenic missense mutation will become more transparent and may be better understood in the context of protein structure. *PTPN11*, associated with Noonan syndrome[106], is predicted, for example, to have three clusters by CLUMP and 3D protein structure analysis reveals that these three clusters define the cleft of the ligand binding site[117] (**Figure 3.4**).

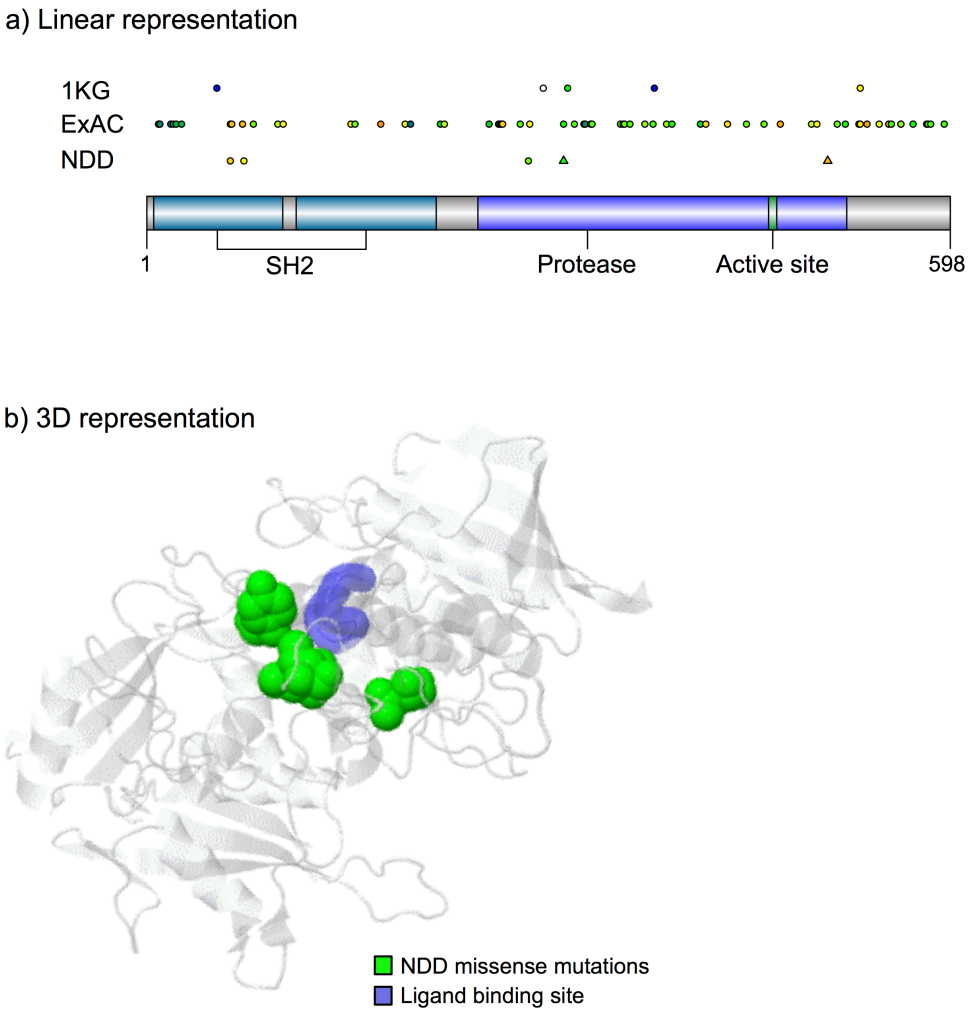

Figure 3.4. Missense mutation clustering in PTPN11. a) Linear representation of the protein (NP_002825.3) with known functional domains annotated. The top two lines show substitutions in controls from 1KG[79] and ExAC[18]. The third line shows *de novo* substitutions in cases in denovo-db[64] v.1.2. The case events fall in three small clusters. **b)** 3D representation of the PTPN11 protein shows that the three clusters of mutations that are far apart in the linear protein result in amino acids changes in close proximity to each other and the ligand binding site after folding[34].

It is interesting that genes associated with hotspots of missense mutation (**Appendix 3.8**) are particularly enriched for presynaptic active zone proteins, FMRP-binding targets, and covalent chromatin modification, although not CHD8 target genes. Accumulating evidence supports a link between the development and function of excitatory synapses in NDD and ASD[95]. Consistent with this, we find 35-fold and 11-fold enrichments of genes regulating postsynaptic membrane potential in genes with a significant burden and significant clustering of *de novo* missense events, respectively. While several scaffolding and intracellular signaling proteins have been associated with ASD and disruption of synaptic function,

including SH3 and multiple ankyrin repeat domain (SHANK) proteins[118], synaptic Ras GTPase-activating (SYNGAP) proteins[119], neurexins[120], neuroligins[121], and others[95], a functional substitution in an essential pore-forming subunit of an excitatory ionotropic glutamate receptor has not been described to our knowledge.

The fact that five patients with phenotypic similarity were identified with a gain-of-function mutation resulting in p.Ala636Thr strongly supports a role for *GRIA1* in ASD and related NDDs. This specific *de novo* missense mutation has been observed before at the homologous position in a highly conserved motif in a different glutamate receptor, GluRδ2[111]. The mutation has a gain-of-function effect, causing constitutive channel opening, neurotoxicity, and degeneration of the cerebellar Purkinje cells in which GluRδ2 is selectively expressed[122]. Both mice and humans with this mutation in GluRδ2 develop ataxia as a direct consequence[112]. This substitution in rodent GluA1 (the protein product of *Gria1*) has the same effect on channel gating[114,115], and here we have replicated this finding in human GluA1. As GluA1 plays an important role in learning and memory[116], there is a biologically plausible link between this *de novo* missense mutation in *GRIA1* and ID.

GRIA1 has been demonstrated to play a key role in early synapse development, with GluA1 homomeric channels being inserted into nascent synapses to provide a calcium-permeable, high-conductance channel prior to being replaced by GluA2-containing channels that mediate long-term synaptic connectivity. Continuing into adulthood, long-term potentiation of excitatory synapses, associated with learning and memory, requires initial insertion of GluA2-absent, calcium permeable AMPA receptors followed by replacement with GluA2-containing receptors[116]. The developmental and adult function of GluA1 in these contexts likely contributes to the ID associated with this substitution. It is interesting to note that loss-of-function of GluA1 in *Gria1* knockout mice leads to impaired synaptic function[123] and behavioral phenotypes, including social behavior deficits and impulsivity[124], which suggests that bidirectional aberration in excitatory signaling can result in similar ASD and NDD phenotypes. Future studies investigating the impact of the mutation resulting in the gain-of-function, Lurcher-like p.Ala636Thr

substitution in synapse development and function will shed additional light on how alterations in excitatory synaptic function contribute to ASD.

Chapter 4: Molecular and phenotypic correlates of *de novo* missense mutation clustering.

This chapter is based on my contributions to two publications:

Coe, B.P., Stessman, H.A.F., Sulovari, A., Geisheker, M.R., Bakken, T.E., Lake, T.E., et al. (2019). Neurodevelopmental disease genes implicated by *de novo* mutations and copy number variation morbidity. *Nature Genetics*. 51(1):106-116. doi: 10.1038/s41588-018-0288-4.

Cogné, B., Ehresmann, S., Beauregard-Lacroix, E., Rousseau, J., Besnard, T., Garcia, T., et al. (2019). Missense variants in the histone acetyltransferase complex component gene *TRRAP* cause autism and syndromic intellectual disability. *American Journal of Human Genetics*. 104(3):530-541. doi: 10.1016/j.ajhg.2019.01.010.

For Coe et al. (2019), I analyzed missense mutations in denovo-db v.1.5 with CLUMP and performed enrichment analyses on the resulting gene set. Using the union set of genes identified in the paper, I performed additional analyses to further interrogate the properties of missense mutations. For Cogné et al. (2019), I identified *TRRAP* mutations in denovo-db, performed validation experiments, and determined inheritance of the events for which we had sample DNA in-house, including a *de novo* missense mutation identified with targeted sequencing. *TRRAP* is an example of a gene that primarily contributes to neurodevelopmental disorders through *de novo* missense mutations.

4.1 Summary

Using two models, we identified a total of 253 genes with a burden of *de novo* mutation, including 145 with excess LGD mutations and 123 with excess missense mutations in patients with NDDs. The genes in this union set are highly interconnected, and the top two modules determined by protein-protein interaction and coexpression are enriched for transcription regulation (Module 1) and neuronal

communication (Module 2). Genes with *de novo* LGD burden share the same enrichment as Module 1 while genes with *de novo* missense burden share the same enrichment as Module 2. Although the two models agree on genes with LGD burden, there is less concordance for genes with missense burden, suggesting different pathways based on mutation type. To better understand the contribution of missense mutations to NDDs, I applied CLUMP, an unsupervised clustering algorithm, and found that 183 genes have *de novo* missense mutations that are significantly more clustered than controls. This set of clustered genes has the same function and expression enrichments as the set of genes with *de novo* missense burden. For one novel gene, *TRRAP*, which recently reached significance with both burden models and CLUMP, phenotypes from 24 patients with *de novo* missense mutations segregate with the location of the substitution, indicating the biological relevance of mutation clustering and highlighting an unknown region of the protein for future functional characterization.

4.2 Introduction

As disease cohorts have grown and sequencing costs have decreased, *de novo* mutations in NDD patients have accumulated in many genes. To differentiate genes with increased mutational load due to gene-specific mutation rate from those that are potentially disease-related, new statistical models have been developed to test for significant enrichment in children with NDD. By taking into account the expected rate of coding *de novo* events, conservation, and other mutational biases, these models make predictions on the expected number of *de novo* events by mutation type in each gene[39,125]. With one of these models[39], we identified 35 genes with a burden of *de novo* missense mutations in NDDs (Chapter 2), and more recently, it was used to identify burden (missense or LGD) in 78 genes[96]. However, more than 400 genes are expected to contribute to NDD pathogenesis under a *de novo* mutation rate model[47].

In this chapter, we take advantage of the growth of ASD and DD/ID cohorts to identify additional genes with burden. To further increase the list of candidate genes, we use a second model[125] in addition to the one described previously[39,126] (Chapter 2). As neither model incorporates characteristics unique to missense mutations, such as their location within the gene, we then use a clustering algorithm[51,126] that

previously identified a set of genes enriched for relevant functions such as neuronal communication (Chapter 3). Finally, we assess the biological relevance of clustered missense mutations in a novel missense-specific gene. Comparison of the characteristics of gene sets identified in different ways, via burden and via clustering, and detailed assessment of genes identified may provide evidence to the validity of clustering as a feature of risk genes.

4.3 Methods

denovo-db: Additional exome and genome studies on individuals with idiopathic NDDs were added to denovo-db (**Table 2.1**) as they were published. Four new studies on ASD[103,127–129] and two new studies on DD/ID[91,130] are included in denovo-db v.1.5 (**Table 4.1**). The new studies from MSSNG[128,129] include mutations that were published in Jiang et al. (2013) and Yuen et al. (2015). Similarly, the data from McRae et al. (2017) includes mutations published in Hurles et al. (2014). For both of these sets of republished data, the older studies were removed to prevent duplicate entries. With the addition of these new studies, the ASD cohort grew from 4,197 cases to 5,624, and the DD/ID cohort nearly tripled from 2,104 cases to 5,303, for a total of 10,927 cases (**Tables 2.1** and **4.1**). The increased sample size allowed us to focus on NDDs that have greater phenotypic overlap, namely ASD and DD/ID.

We annotated all *de novo* SNV and indel variants to RefSeq transcripts in GRCh37/hg19 using SnpEff. When multiple transcripts were present, we chose the one with the most severe variants. This dataset includes 2,357 LGD mutations and 9,815 missense mutations (12,172 *de novo* mutations (DNMs) total) (**Table 4.1**). When necessary, gene symbols were adjusted to match those used in the individual analysis models. If no model was generated for a gene of interest, “no model” is indicated in the significance column (**Appendix 4.1**).

Table 4.1. NDD cohorts and *de novo* mutations in denovo-db v.1.5.

Study	Study diagnosis	Cases	LGD DNMs	Missense DNMs
SSC[25,28,39,61,85]	ASD	2,508	449	2,007
ASC[75]	ASD	1,445	189	1,074
MSSNG[128,129]	ASD	1,625	267	1,307
NIMH[65]	ASD	10	2	8
Lee2014[62]	ASD	1	0	2
Tavassoli2014[67]	ASD	1	1	1
Hashimoto2015[56]	ASD	30	3	27
Moreno-Ramos2015[103]	ASD	4	0	2
DDD2017[91]	DD	4,293	1,198	4,624
deLigt2012[74]	ID	100	12	48
Rauch2012[66]	ID	51	23	56
Halvardson2016[130]	ID	39	4	20
Lelieveld2016[63]	ID	820	209	639
	TOTAL	10,927	2,357	9,815

Gene burden: Two statistical models were used to calculate enrichment of LGD and missense DNMs per gene. Both models use estimations of mutation rate to generate prior probabilities for observing a specific number and class of mutations for a given gene. The first model incorporates locus-specific transition/transversion/indel rates and chimpanzee–human coding sequence divergence to estimate the number of expected DNMs[39]. The second model, denovolyzeR, estimates mutation rates from trinucleotide context and macaque–human comparisons over a +/-1 Mbp window[125]. It also incorporates exome read depth and accommodates known mutational biases such as CpG hotspots. We also used a modified version of the chimpanzee–human (CH) model[131] which incorporates CADD scores[90]. This allowed us to test for enrichment of missense DNMs with CADD Phred scores over 30 (MIS30), which are predicted to be in the most severe 0.1% of mutations. MIS30 mutations are more likely to have functional effects as severe as LGD mutations and are significantly enriched in NDD cases compared to controls[126].

Both models were run using the default settings, and a baseline rate of 1.8 *de novo* variants per individual was used for the CH model. For each test (missense, MIS30, LGD) in both models, we calculated a q-value by the Benjamini-Hochberg procedure based on the number of genes in the model (18,946 for CH and 19,618 for denovolyzeR). Based on expectations that 50% of DNMs are likely related to disease, genes with a q-value less than 0.1 (representing a false discovery rate (FDR) near 5%) and 2 or more

DNMs were considered[24,102]. If a gene was found to have a significant association with NDDs in recent large-scale sequencing studies[49,75,91,96,129] or curated databases (OMIM [http://omim.org/], ID Gene Database Project [http://gfuncpathdb.ucdenver.edu/iddrc/iddrc/home.php], and SFARI Gene [https://gene.sfari.org/]), it was not considered as novel. SFARI genes of any score category were also not considered to be novel. A literature search for case reports or human studies of NDDs (ASD, ID, DD, or mental retardation) found evidence for 31 additional genes. The remaining 49 genes, with no published findings, were considered novel.

Network and enrichment analysis: Clustered gene modules were identified with the MAGI (merging affected genes into integrated networks) enrichment tool[132] with default settings. We used coexpression and physical interaction data from geneMANIA[133] to further assess and visualize the results. Module significance was assessed by permuting DNMs 100 times across genes according to their mutation rates in the CH model[39]. Gene ontology enrichment was tested with PANTHER (database 2019-03-12) using the GO slim biological process annotation. Two-tailed binomial tests were performed and Bonferroni-corrected p-values (p_{adj}) are reported. We also tested gene sets for enrichment of expression in different brain cell types with Cell-type Specific Expression Analysis[134] (CSEA) and across different brain regions during development with Tissue Specific Expression Analysis[134] (TSEA). Cell type data come from mice and brain region data come from human (BrainSpan: Atlas of the Developing Human Brain [http://www.brainspan.org/]). Expression enrichments were tested with one-tailed Fisher's exact tests, and Benjamini-Hochberg corrected p-values are reported as q-values.

CLUMP: We implemented the permutation (-z 1000) and minimum mutation options (-m 2) and calculated a p-value based on the null distribution of case and control CLUMP score differences. To enable exact transcript comparisons between cases and controls, all variants were re-annotated with CRAVAT[105] resulting in 9,772 *de novo* missense variants in cases. The control set included private missense mutations in individuals from ExAC[93] without neuropsychiatric disorders (N = 45,376; 1,466,439 mutations). All genes with at least two mutations in cases and controls were tested (N = 1,930). For each gene we initially performed 1,000 simulations. In genes where nominal significance was reached, we

performed 1,000,000 simulations and calculated an empirical p-value and q-value (Benjamini-Hochberg corrected).

***TRRAP*:** *De novo* mutations in *TRRAP* were identified in individuals with NDDs through an international collaboration and from the web-based tool GeneMatcher [https://www.genematcher.org/]. The subjects were part of research and clinical cohorts that underwent whole-exome and targeted sequencing. Clinicians associated with each site provided phenotypic information for evaluation. All patients gave consent for research procedures. We performed Sanger sequencing on variants in samples from SSC[21] to test call validity and inheritance. Additional variants were identified in denovo-db[79] v.1.5 and via targeted sequencing with smMIPs on 17,688 individuals with idiopathic NDDs who had not been exome sequenced[126]. Probes for *TRRAP* were designed to target residue p.Trp1866 (NP_001231509.1), at which two cases in denovo-db v.1.5 have a *de novo* substitution. These two events were assessed in Cogné et al. (2019). Evidence for mutation pathogenicity was determined in three ways: 1) *de novo* inheritance; 2) lack of detection in controls (gnomAD[93] v.2.1, individuals without neuropsychiatric diagnoses; exomes, N = 104,068; genomes, N = 10,636; total N = 114,704); and 3) location in a conserved region of the protein. Mutation distribution was assessed with two-sided Fisher's exact tests. *De novo* burden was assessed with a modified version of the CH model[39] that calculates the relative mutation rate for a selected region of a gene and normalizes by gene-specific human–chimpanzee divergence[126] (Chapter 2). We assumed a mutation rate of 1.8 *de novo* coding variants per generation and applied a one-tailed binomial test to assess burden. We used a conservative Bonferroni family-wise error rate (FWER) correction based on the number of same-size windows in the genome and report the adjusted p-value (p_{adj}).

4.4 Results

Genes enriched for *de novo* SNV mutation. Combined, the two models (union set) identify 253 candidate NDD genes with at least two mutations of one type (LGD, missense, or MIS30) and a burden of DNM at a false discovery rate (FDR) <5% (**Tables 4.2** and **4.3**). With the more stringent Bonferroni

FWER correction, the union includes 124 genes. This 253 candidate gene set consists of 145 genes with a burden of LGD mutations and 123 genes with a burden of missense mutations (**Figure 4.1a**). Twenty-nine of these candidate genes in the union set have an excess of both LGD and missense mutations. For genes without LGD burden, 94 have a burden of missense mutations and an additional 14 have a burden of MIS30 mutations. The two models implicate similar genes with LGD burden, with 72.4% (105/145) of genes identified by both models (**Figure 4.1b**). For missense mutations, however, only 51.2% (63/123) of genes are identified by both models. Among the union genes, 82.6% (209/253) have no detected LGD or missense DNM in controls in denovo-db v.1.5 (N = 2,278). None of the genes with recurrent mutations in controls are significant after exome-wide FDR correction.

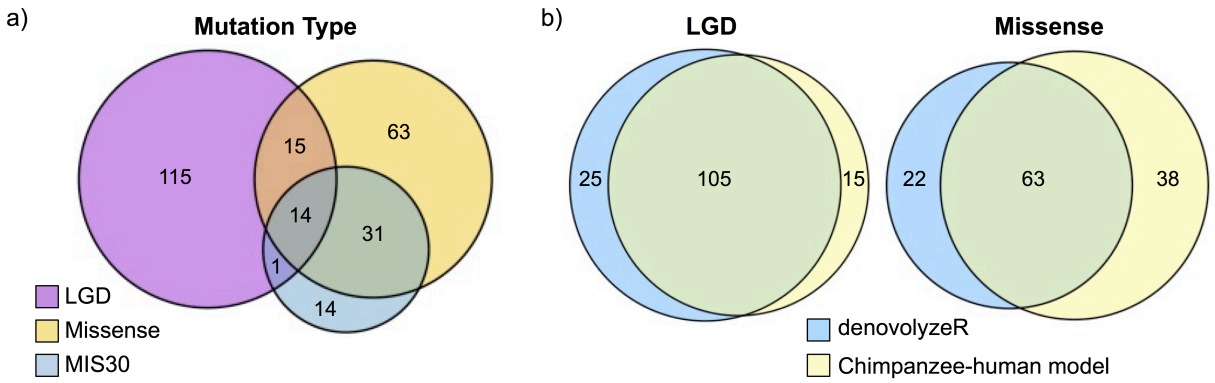

Figure 4.1. Genes identified with burden of de novo mutations using two models. a) Types of mutation burden identified in individuals with ASD, DD, or ID (N = 10,927) by two models. Of the 253 total genes identified, 115 have only a burden of LGD mutations, 63 have only a burden of missense mutations, and 14 have only a burden of missense mutations with CADD[90] Phred score over 30 (MIS30). Twenty-nine genes have an excess of LGD and missense mutations, and one additional gene has an excess of both LGD and MIS30 mutations. **b)** Comparison of genes identified with burden by two models shows extensive overlap. For LGD mutations, denovolyzeR[125] finds 130 genes with a burden and the CH model[39] finds 120 genes; 105/145 (72.4%) genes with LGD burden are identified by both models. For missense mutations, 85 genes are identified with denovolyzeR and 101 with the CH model. Only 63/123 (51.2%) genes have a significant missense burden in both models.

Table 4.2. Number of genes identified with *de novo* mutation burden in 10,927 NDD cases.

Burden type	denovolyzeR	CH model	Union (FDR (FWER))	Intersection (FDR (FWER))
LGD	108	92	115 (62)	85 (49)
Missense	63	46	63 (20)	42 (22)
MIS30	NA	14	14 (10)	NA
LGD and missense	22	13	15 (6)	21 (8)
LGD and MIS30	NA	1	1 (3)	NA
Missense and MIS30	NA	28	31 (18)	NA
All	NA	14	14 (5)	NA
TOTAL	193	208	253 (124)	148 (79)

Adapted from Coe et al. (2019). Counts are number of genes with a q-value < 0.05 (FDR corrected). Count of genes with Bonferroni FWER corrected p-value < $5x10^{-7}$ are in brackets.

Table 4.3. Union set of 253 genes identified by denovolyzeR or the CH model as having *de novo* mutation burden (FDR < 5%) in 10,927 individuals with NDDs. From Coe et al. (2019).

Burden type	Genes
LGD	*ADNP*[a*], *AHDC1*[a*], *ANK2*[a], *ANKRD11*[a*], *ANP32A*[*], *ARID1*[a*], *ARID2*[a], *ASH1L*[a*], *ASXL1*[a], *ASXL3*[a*], *AUTS2*[a], *BCL11A*[a*], *BRPF1*[a], *CAPRIN1*[a*], *CASZ1*[a], *CDC42BPB*[a], *CDKL5*[a], *CHAMP1*[a*], *CHD7*[a], *CHD8*[a*], *CLTC*[a], *CNKSR2*[a*], *CNOT3*[a], *CTNNB1*[a*], *CUL3*[a*], *DLG4*[a*], *DSCAM*[a], *DVL3*[a], *EBF3*[a], *EHMT1*[a*], *ENO3*[a*], *EP300*[a*], *FAM200A*[a], *FAM200B*[a], *FOSL2*, *FOXP2*[a], *GATAD2B*[a*], *HIST1H1E*[a], *HIVEP2*[a], *HIVEP3*[a*], *HNRNPD*[a], *IRF2BPL*[a], *KANSL1*[a*], *KAT6A*[a*], *KAT6B*[a*], *KCNS3*[a], *KDM5B*[a*], *KDM6A*[a], *KDM6B*[a], *KIAA2022*[a*], *KIF11*[a], *KMT2A*[a*], *KMT2C*[a], *LARP4B*[a], *LEO1*[a*], *MBD5*[a*], *MEIS2*[a*], *MSL3*[a*], *NAA15*[a*], *NFE2L3*[a], *NONO*[a*], *NSD1*[a*], *ODC1*[a], *PDHA1*[a], *PHF12*[a], *PHF21A*[a], PHF3[a], PHIP[a], POU3F3[a], PPM1D[a*], PRR12[a], PTCHD1[a], QRICH1[a*], RAI1[a], RPL26[a], SET[a*], SETBP1[a*], SETD2[a], SETD5[a*], SHANK3[a*], SIN3A[a], SKIDA1[a], SMC1A[a*], SON[a], SOX5[a*], SPAST[a], SPEN[a], SPRY2[a], SRCAP[a*], SRRM2, SRSF11[a], STARD9[a], SUV420H1[a*], SYNCRIP[a*], SYNGAP1[a*], TAB2[a], TBR1[a], TCF12[a], TCF20[a*], TNRC6B, TRA2B[a], TRIP12[a*], UPF3B[a*], USP9X[a*], VEZF1, WAC[a*], WDFY3[a], WDR45[a*], WDR87[a], WHSC1[a], YTHDF3[a], ZBTB18[a*], ZBTB7A[a], ZC4H2[a], ZNF292[a]
LGD and missense	*CHD2*[a*], *CREBBP*[a*], *DYRK1A*[a*], *FBXO11*[a], *FOXG1*[a*], *FOXP1*[a*], *HNRNPU*[a*], *MEF2C*[a*], *MYT1L*[a*], *NFIX*[a*], *POGZ*[a*], *PTEN*[a*], *PURA*[a*], *TLK2*[a*], *WDR26*[a*]
LGD and MIS30	*TCF7L2*[a]
LGD, missense, and MIS30	*CASK*[a*], *DDX3X*[a*], *HDAC8*[a*], *IQSEC2*[a*], *MECP2*[a*], *MED13L*[a*], *PPP2R5D*[a*], *PUF60*[a*], *SATB2*[a*], *SCN2A*[a*], *SLC6A1*[a*], *STXBP1*[a*], *TBL1XR1*[a*], *TCF4*[a*]
Missense	*ABI2*[a], *ACHE*[a], *ADAP1*[a*], *AGAP2*[a], *AGO1*[a], *AGO4*[a*], *AQP10*[a], *BRAF*[a*], *BTF3*[a], *C2orf42*[a], *CABP7*[a], *CAPN15*[a*], *CBL*[a], *CHD4*[a], *CLASP1*[a], *DEAF1*[a], *DLX3*[a*], *DNM1*[a], *EGLN2*[a], *GABRB2*[a*], *GABRB3*[a], *GLRA2*[a], *GNAI1*[a*], *HMGXB3*[a], *HUWE1*[a], *ITPR1*[a*], *KCNC1*[a], *KCNJ6*[a], *MAPK3*[a], *MTF2*[a], *MYO1E*[a], *PBX1*[a], *PLAC8L1*[a], *PLK5*[a], *PPP1CB*[a*], *PRKCA*[a], *PRKD1*[a], *PRPF18*[a], *PSMG4*[a], *PTPN11*[a*], *RAC1*[a*], *RFX8*[a], *RRP8*[a], *RYR2*[a], *SETD1B*[a], *SF3B1*[a], *SHISA6*[a], *SMAD4*[a], *SMARCD1*[a*], *SMC3*[a], *SNAPC5*[a*], *SNX5*[a*], *SUSD4*[a], *SYT1*[a], *TAOK1*[a], *TMEM178A*[a*], *TMEM42*[a], *TNPO3*[a], *TRAF7*[a], ***TRRAP***[a], *UNC80*[a], *VAMP2*[a], *WDFY4*[a], *YWHAG*[a]
MIS30	*ACTC1*[a], *AGO3*[a], *CACNA1E*[a], *FAM104A*[a], *HIST1H2AC*[a], *KIF5C*[a], *PACS2*[a*], *PAPOLG*[a], *PDK2*[a], *SEPT10*[a], *STC1*[a], *TAF1*[a*], *TNPO2*[a*], *U2AF2*[a]
Missense and MIS30	*CDK13*[a*], *CHD3*[a*], *COL4A3BP*[a*], *CSNK2A1*[a*], *CTCF*[a*], *DNMT3A*[a*], *DYNC1H1*[a*], *EEF1A2*[a*], *EFTUD2*[a*], *GNAO1*[a*], *GRIN2B*[a*], *HECW2*[a*], *KCND3*[a], *KCNH1*[a*], *KCNQ2*[a*], *KCNQ3*[a*], *KIF1A*[a*], *MAP2K1*[a*], *NAA10*[a*], *NR2F1*[a], *NR4A2*[a], *PACS1*[a*], *PIK3CA*[a*], *PPP2R1A*[a*], *RAB11A*[a], *SCN8A*[a*], *SMARCA2*[a*], *SMARCA4*[a*], *TRIO*[a*], *ZMYND11*[a*]

[a]Gene also in the intersection set.

*Gene in FWER exome-wide significance ($p < 5x10^{-7}$) set.

For 204/253 genes in the union set, additional evidence for association with NDDs exists in the literature. Of the 49 novel genes, 15 have a burden of LGD mutations, while 26 and eight have a burden of missense and MIS30 mutations, respectively (**Appendix 4.1**). Of the novel missense genes, only four have LGD mutations in cases in denovo-db v.1.5. The remaining 22 are missense-specific genes. This may indicate that LGD mutations in these genes lead to a phenotypic outcome not covered in the studies or that these genes are so critical that loss of function is incompatible with life. The larger number of novel missense-specific genes relative to novel LGD genes points to the increased focus that has been placed on LGD mutations, as well as the power of large sample sizes to identify pathogenic missense mutations. This is especially important given the high rate of incidental missense events, as evidenced by the missense rate in unaffected controls (**Figure 2.1**).

Network enrichment and patterns of brain expression. The union set of 253 genes is strongly enriched for genes with related functions, consistent with previous observations[28,49,72,75,84,126,132,135–138]. Preliminary analysis with the STRING database[139] finds a highly significant 1.8-fold enrichment (1,067 edges vs. 573 expected) in protein interactions ($p < 1.0 \times 10^{-16}$, one-tailed hypergeometric test). To further characterize these interactions, we used MAGI[132], an enrichment tool that identifies biological subnetworks based on protein-protein interactions (PPI) and gene coexpression profiles during brain development. PANTHER enrichment analysis of genes in the three top modules found with MAGI ($p < 0.01$, one-tailed permutation test) recapitulates known functions of NDD risk genes (**Figure 4.2** and **Appendix 4.2**).

Module 1, which includes 20 genes, is characterized by 'regulation of transcription by RNA polymerase II' (8 observed vs. 0.62 expected, 9.68-fold enrichment, $p_{adj} = 3.44 \times 10^{-2}$, two-sided binomial test). Module 2 contains 20 genes that are enriched for two functions: 1) 'protein-containing complex localization' (3 observed vs. 0.02 expected, >100-fold enrichment, $p_{adj} = 1.88 \times 10^{-3}$, two-sided binomial test) and 2) 'chemical synaptic transmission' (6 observed vs. 0.36 expected, 16.60-fold enrichment, $p_{adj} = 1.76 \times 10^{-3}$, two-sided binomial test). Finally, the 35 genes in module 3 are enriched for three functions: 1) 'protein ubiquitination' (6 observed vs. 0.34 expected, 17.50-fold enrichment, $p_{adj} = 1.66 \times 10^{-3}$, two-sided binomial

test), 2) ‘proteolysis involved in cellular protein catabolic process’ (7 observed vs. 0.43 expected, 16.14-fold enrichment, p_{adj} = 3.30x10^{-4}, two-sided binomial test), and 3) ‘regulation of metabolic process’ (11 observed vs. 2.31 expected, 18.15-fold enrichment, p_{adj} = 1.47x10^{-2}, two-sided binomial test).

Although the first two modules identified with MAGI are both composed of genes with varying types of mutational burden, the modules may point to the differing roles of LGD and missense mutations in disease. The same PANTHER analysis of 145 genes with LGD burden shows enrichment only for transcription, like Module 1: ‘positive regulation of transcription by RNA polymerase II’ (13 observed vs. 1.80 expected, 7.24-fold enrichment, p_{adj} = 5.92x10^{-5}, two-sided binomial test) and ‘transcription by RNA polymerase II’ (28 observed vs. 6.98 expected, 4.01-fold enrichment, p_{adj} = 4.89x10^{-7}, two-sided binomial test). And the 123 genes with missense burden show enrichment for aspects of neuronal communication, similar to Module 2: ‘regulation of membrane potential’ (8 observed vs. 0.85 expected, 9.44-fold enrichment, p_{adj} = 3.79x10^{-3}, two-sided binomial test) and ‘chemical synaptic transmission’ (11 observed vs. 2.01 expected, 5.47-fold enrichment, p_{adj} = 9.47x10^{-3}, two-sided binomial test).

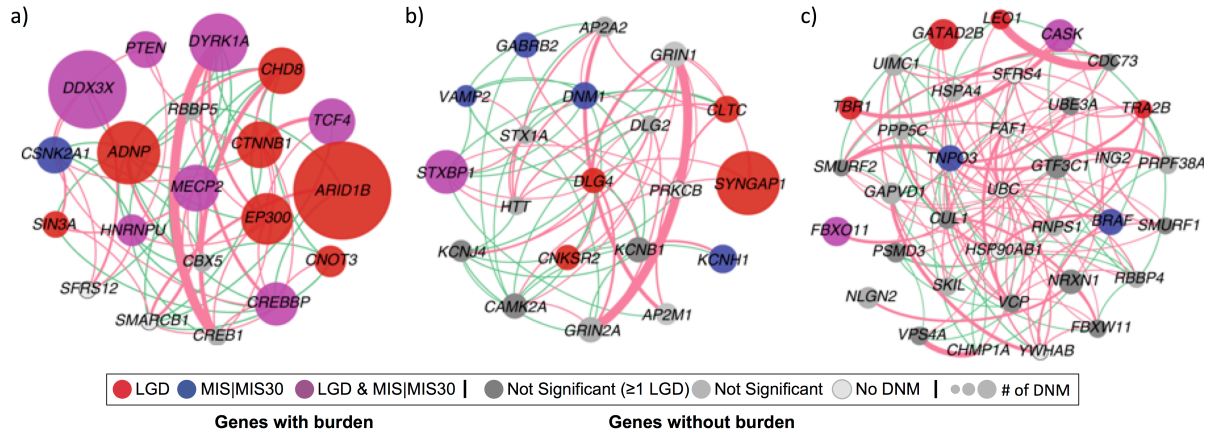

Figure 4.2. Modules identified by MAGI analysis of 253 genes with burden of mutations. Adapted from Coe et al. (2019)**. a)** Module 1 contains 20 genes, seven with LGD burden only, one with missense burden only, seven with both types of burden, and five with no burden. This module is enriched for transcription regulation (8 observed vs. 0.62 expected, 9.68-fold enrichment, p_{adj} = 3.44x10^{-2}, two-sided binomial test). **b)** Module 2 also contains 20 genes. Four have LGD burden alone, four have missense burden alone, and one has both types. Of the 11 with no burden, only three genes have LGD mutations in denovo-db v.1.5, while all but one have missense mutations. This module is enriched for localization of complexes (3 observed vs. 0.02 expected, >100-fold enrichment, p_{adj} = 1.88x10^{-3}, two-sided binomial test) and synaptic transmission (6 observed vs. 0.36 expected, 16.60-fold enrichment, p_{adj} = 1.76x10^{-3}, two-sided binomial test). **c)** Module 3, with 35 genes, contains four with LGD burden, two with missense burden, and two with both types. It is enriched for diverse functions including protein ubiquitination (6 observed vs. 0.34 expected, 17.50-fold enrichment, p_{adj} = 1.66x10^{-3}, two-sided binomial test) and regulation of metabolic processes (11 observed vs. 2.31 expected, 18.15-fold enrichment, p_{adj} = 1.47x10^{-2}, two-sided binomial test).

In addition to enrichments in PPI and gene function, we also noted enrichments in gene expression. The union set of 253 genes shows enrichment for expression in most regions of the brain during fetal development, with the strongest enrichment for cortical expression (**Figure 4.3a**). A similar pattern of expression is seen in genes with LGD burden (N = 145; **Appendix 4.3**), but the pattern shifts when looking at genes with missense burden (N = 123; **Figure 4.3b**). For these genes, enrichment of expression in the fetal period is limited to the amygdala and cortex. Looking more closely at the level of neuron subtypes during development, the union set is enriched in striatal medium spiny neurons that express dopaminergic receptors D1 and D2 ($q = 0.013$ and $q = 0.011$, one-tailed Fisher's exact test, Benjamini-Hochberg (BH) correction) at a pSI (specificity index P value) threshold of 0.05. The union set, as well as the 145 genes with LGD burden, are also strongly enriched for expression in retinal rods ($q = 0.013$ (union) and $q = 6.42 \times 10^{-5}$ (LGD), one-tailed Fisher's exact test, BH correction). For the 123 genes with missense burden, enrichments remain in D1- and D2-positive neurons ($q = 0.034$ (both), one-tailed Fisher's exact test, BH correction) but there is no enrichment in rods ($q = 1$, one-tailed Fisher's exact test, BH correction).

Clustered missense mutations. Given our observation of a larger number of novel missense-specific genes despite a larger total number of LGD-significant genes, and the increased predicted rate of missense mutations in NDD pathology[28], we hypothesize that many important missense-specific genes have not yet been identified. We therefore assessed the clustering of *de novo* missense mutations, as this pattern of mutation is associated with disease. Using CLUMP, an unsupervised clustering algorithm, we compared *de novo* missense mutations in cases in denovo-db v.1.5 with private missense mutations in the ExAC[93] control database (N = 45,376; 1,466,439 mutations) and found that 183 genes have *de novo* missense mutations that are significantly more clustered (1×10^6 permutations; $p < 0.05$) in cases than controls (**Appendix 4.4**). After correcting for FDR, significant clustering ($q < 0.1$) remains for 102 genes.

Only 14.8% (27/183) of genes with significant clustering have an excess of missense DNMs in the burden analyses (union set) (**Appendix 4.4**). However, although largely distinct from the set of genes identified

with whole-gene burden, this set of clustered genes shows similar enrichments. Assessment of gene ontology finds that clustered genes are enriched for roles in membrane potential regulation (9 observed vs. 1.25 expected, 7.19-fold enrichment, $p_{adj} = 8.52x10^{-3}$), but not functions related to gene transcription, as is seen in genes with LGD burden. During fetal development, this set of genes is enriched for expression in the amygdala and cortex (**Figure 4.3c**) and in striatal D2-positive medium spiny neurons (nominal p = 0.005, q = 0.177). These genes, especially those expressed during brain development, should be considered as candidates for further follow-up studies.

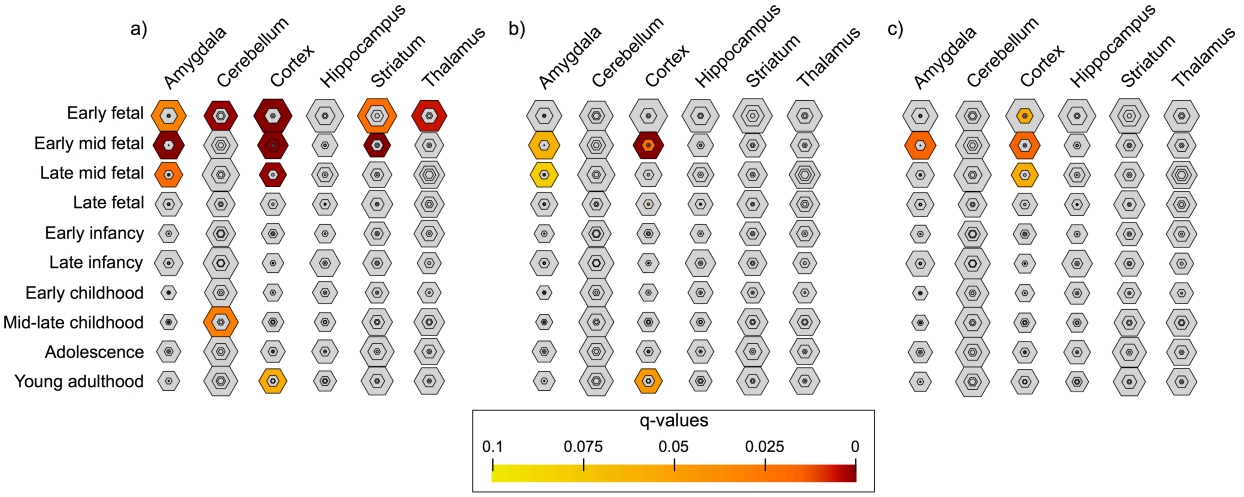

Figure 4.3. Tissue-specific Enrichment Analysis shows enrichment for genes expressed in different brain regions during development. a) Genes in the union set (N = 253) are expressed broadly across the brain during fetal development but have a more limited expression in childhood (cerebellum) and adulthood (cortex). Color represents one-tailed Fisher's exact p-values with Benjamini-Hochberg correction (q-values), with regions of shading closer to the center indicating greater tissue sensitivity (see graphic legend). **b)** For the 123/235 genes with missense burden, expression is limited to the amygdala and cortex during fetal development. Cortical expression is also seen during adulthood. **c)** Genes (N = 183) with missense mutations that are significantly (p < 0.05) more clustered in cases (N = 10,927) than ExAC controls (N = 45,376) also show enrichment for expression in amygdala and cortex during fetal development.

TRRAP: With both *de novo* burden models and clustering analysis, we identified a novel missense-specific risk gene, *TRRAP* (**Table 4.3**). This gene has been linked to cancer, but its role in NDDs had not been previously described. It is of particular interest because two mutations in denovo-db[79] are missense and occur at the same highly conserved residue, p.Trp1866 (NP_001231509.1), in unrelated individuals (Chapter 2.4). To further assess the potential role of this gene in NDDs, we used targeted sequencing on

17,688 individuals with idiopathic NDDs (Chapter 3). We identified nine additional missense variants clustered around the recurrent site. One of these substitutions, p.Gly1883Arg, was found to be *de novo* and four are inherited (**Appendix 4.5**). Inheritance of the other four variants is unknown due to limited availability of parental DNA.

We collaborated with several other groups that identified *TRRAP* mutations in individuals with NDDs[140]. In addition to five mutations from denovo-db (three with ASD (SSC[21]) and two with DD (DDD[76])) and the *de novo* event identified with targeted sequencing (Chapter 3.4), variants of interest were found in 18 other individuals (**Figure 4.4a**). Notably, all variants are missense. All but one of the 24 total missense events are *de novo*; p.Glu1106Lys, seen in two sisters, was inherited from a mother with low-level mosaicism. None of the events have been seen in controls in gnomAD[93] v.2.1 (N = 114,704) and they all fall within regions that are depleted for missense variation in these controls. All the variants are predicted to be deleterious by CADD[90] (Phred score > 20) and pathogenic by SIFT[141] and polyphen2 hvar[142]. Further, these substitutions all occur at residues that are highly conserved (**Figure 4.4b**), especially among vertebrates. In addition to the two substitutions at p.Trp1866 in denovo-db[79], three of the substitutions were seen in multiple individuals: p.Ala1043Thr was identified in five individuals and p.Glu1106Lys and p.Gly1883Arg were identified in two individuals each. Nine of the events, including p.Ala1043Thr, occur at CpG sites, known to be highly mutable.

Substitutions in TRRAP in these 24 individuals can be separated into two clusters – Cluster 1, from p.Ile1031 to p.Gly1159, and Cluster 2, from p.Arg1859 to p.Pro1932 – with additional surrounding events (**Figure 4.4a**). Across the gene, a significantly higher fraction of mutations falls in one of these two clusters in cases versus private missense events in gnomAD[93] controls (Cluster 1, $p = 4.26 \times 10^{-11}$, OR = 23.1 [9.2-57.8]; Cluster 2, $p = 6.57 \times 10^{-7}$, OR = 23.9 [6.9-79.6]; two-tailed Fisher's exact test). Further, there is a burden of *de novo* substitutions in both of these clusters (Cluster 1, $p_{adj} = 5.52 \times 10^{-14}$; Cluster 2, $p_{adj} = 7.22 \times 10^{-6}$). Since publication of Cogné et al. (2019), an additional *de novo* substitution in Cluster 1, p.Val1113Met, has been identified in an individual with DD[79,91]. And the three additional missense events in four individuals identified with targeted sequencing (Chapter 3.4) are in Cluster 2 (**Figure 4.4a**). The

Cluster 1 substitution and the substitution seen recurrently in Cluster 2, p.Asn1851Lys, are not present in gnomAD controls[93], have CADD scores that suggest deleteriousness[90] (Phred scores 23.1 and 32, respectively), and impact highly conserved residues (**Figure 4.4b**). These additional findings strengthen the association of these two regions with disease risk.

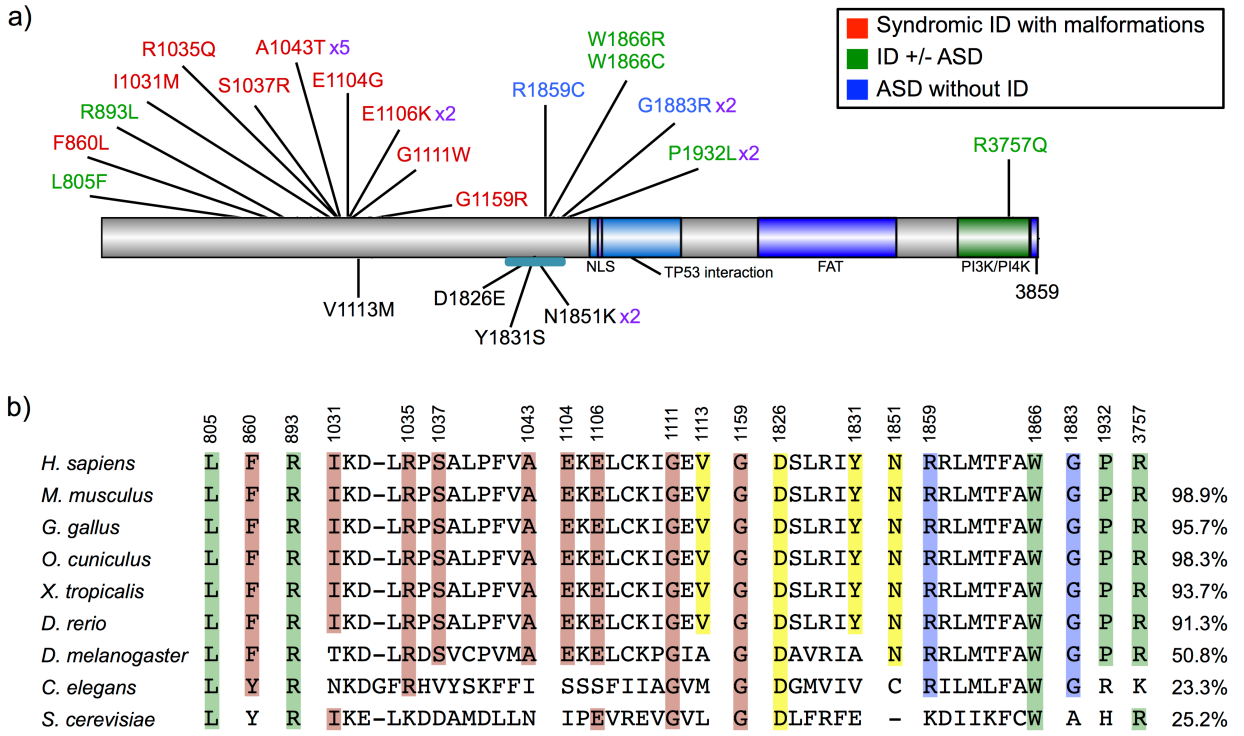

Figure 4.4. Substitutions in TRRAP cluster in two conserved regions and show genotype–phenotype correlation. a) Linear representation of TRRAP (NP_001231509.1). Colored regions on protein indicate known functional domains. NLS, nuclear localization signal. FAT, FRAP-ATM-TRRAP. PI3K/PI4K, phosphatidylinositol 3 kinase/phosphatidylinositol 4 kinase. FATC, FAT C-terminal. Substitutions above the protein are described in Cogné et al. (2019) and patient phenotype, indicated by text color, is well characterized. These substitutions also have evidence of pathogenicity (CADD Phred score > 20, not present in gnomAD controls, and at a conserved residue). The substitutions fall in two clusters, Cluster 1 (p.Ile1031 to p.Gly1159) and Cluster 2 (p.Arg1859 to p.Pro1932), with additional substitutions surrounding. Individuals with substitutions in Cluster 1 have a more severe phenotype, with severe ID and physical malformations (**Table 4.4**). Individuals with substitutions in Cluster 2 and out of the clusters have ID and/or ASD. Substitutions below the protein were identified in denovo-db[79] v.1.5 and with targeted sequencing[126]. Detailed phenotypic information is not available but the broad diagnoses for these individuals is in accordance with the clusters in which their events reside, i.e., the individual with the substitution in Cluster 1 has DD and the individuals with substitutions in Cluster 2 have ASD. **b)** Multiple sequence alignment of TRRAP orthologs across different species. Highlight color indicates the individual phenotype, as above, with yellow representing substitutions not described in Cogné et al. (2019). Percent identify with human TRRAP is shown on the right for each species.

Assessment of patient phenotypes shows correlation with genotype, with features segregating according to mutation clustering (**Figure 4.4** and **Table 4.4**). The 13 patients with substitutions in Cluster 1 all have global DD/ID, severe in most cases. They also all have physical malformations, including microcephaly (6/13) and other brain abnormalities, cleft lip/palate (5/13), and heart (7/10 tested), abdominal (7/10 tested), and urogenital (5/13) anomalies. This may indicate that Cluster 1 is part of an uncharacterized functional domain.

Conversely, the seven patients with substitutions in Cluster 2 and the four with substitutions elsewhere in the protein have a milder phenotype, largely limited to ASD and/or ID. Five out of eleven have ASD and an additional three display characteristics of ASD but have not been formally diagnosed. Three individuals had delayed speech without ID (IQ > 70), and all remaining individuals had both DD and mild to severe ID. Additionally, four individuals have epilepsy. Most did not have malformations, although the patient with the variant causing p.Phe860Leu has microcephaly and heart defects, and the patient with the p.Trp1866Arg substitution has lacrimal duct aplasia and optic disc colobomas. Additional correlation between phenotype and genotype can be seen in the cases with recurrent mutations. Both individuals with an amino acid substitution at p.Trp1866 have severe ID and epilepsy, both sisters with the p.Pro1932Leu substitution are severely affected, and both individuals with p.Gly1883Arg have IQ >70. Although detailed phenotypic information is not available for the five additional variants found, the broad diagnoses of the individuals are concordant – the Cluster 1 individual has DD and the Cluster 2 individuals have ASD (**Figure 4.4b**). This segregation of patient phenotypes based on substitution location, especially within Cluster 1, points to varying functional roles across the protein, although both clusters occur in regions that do not yet have functions attributed.

Table 4.4. Phenotypes of individuals with missense mutations in *TRRAP*.

		p.(Leu805Phe)	p.(Phe860Leu)	p.(Arg893Leu)	p.(Ile1031Met)	p.(Arg1035Gln)	p.(Ser1037Arg)	p.(Ala1043Thr)	p.(Ala1043Thr)	p.(Ala1043Thr)	p.(Ala1043Thr)	p.(Ala1043Thr)	p.(Glu1104Gly)	p.(Glu1106Lys)	p.(Glu1106Lys)	p.(Gly1111Trp)	p.(Gly1159Arg)	p.(Arg1859Cys)	p.(Trp1866Arg)	p.(Trp1866Cys)	p.(Gly1883Arg)	p.(Gly1883Arg)	p.(Pro1932Leu)	p.(Pro1932Leu)	p.(Arg3757Gln)
Gender		F	F	F	F	F	F	F	F	M	M	M	M	F	M	M	F	M	M	F	F	M	F	F	F
Age at examination (years)		8	3	4	4.5	8.5	29	1 day	4	11	4	8	5.5	14	12.5	12	2	8.5	14	8.5	10	11	24	19	22
NDDs	**Global DD**	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	ID	Sev	NA	No	Sev	Mild	Yes	NA	NA	Sev	Sev	Sev	Sev	Yes	Sev	Yes	Yes	No	Mod-sev	Sev	No	No	Sev	Sev	Yes
	ASD	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	-	-	-	-
Physical malformations	**Microcephaly (<-2.5 SD)**	-	+	-	+	-	+	+	+	-	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-
	Short stature	-	+	-	+	-	+	-	+	-	-	NA	+	-	-	-	-	-	+	+	-	-	-	-	-
	Cleft lip/palate	-	-	-	-	-	-	+	+	-	+	+	-	+	-	-	-	-	-	-	-	-	-	-	-
	Facial dysmorphology	+	+	-	+	+	+	+	-	+	+	+	+	+	+	-	+	-	+	+	+	+	+	+	-
	Cerebellar hypoplasia	-	NA	-	+	NA	-	+	NA	-	+	+	-	-	-	-	+	NA	-	-	NA	-	-	NA	-
	Cerebral abnormalities	-	NA	-	+	NA	+	+	NA	+	+	+	-	-	-	-	-	NA	-	-	NA	-	-	NA	-
	Cardiac malformations	-	+	-	-	+	-	+	+	+	NA	+	+	+	+	+	-	NA	NA	NA	NA	NA	NA	NA	NA
	Renal malformations	-	NA	NA	-	+	-	-	+	-	-	+	-	-	-	+	+	NA	-	-	NA	-	NA	NA	NA
	Genital malformations	-	-	-	-	-	-	-	+	+	+	+	-	-	-	-	+	-	-	-	-	-	-	-	-
	Scoliosis	-	-	-	-	-	-	-	-	+	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-
	Dysplastic nails	-	-	-	+	+	-	-	-	-	+	+	-	+	+	-	+	-	-	-	-	-	-	-	-
	Accessory nipple	-	-	-	+	-	-	-	-	-	+	-	-	-	+	-	-	-	-	-	-	+	-	-	-
Neurological	**Hypotonia**	-	+	+	+	-	-	-	-	+	+	+	-	-	-	-	-	-	-	+	-	+	-	-	-
	Lower limb hyperreflexia	+	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	+	-	-	-	+	-	+
	Feeding difficulties	-	+	-	+	-	-	-	+	-	+	+	-	+	-	+	-	-	-	-	-	-	-	-	-
	Hearing impairment	-	-	-	+	-	-	-	+	-	-	-	-	-	-	-	+	-	-	-	-	-	-		-
	Visual impairment	-	-	-	-	-	-	+	+	+	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-
	Seizures	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	+	+	+	-	-	-	+

Mutations annotated to grch37 NP_001231509.1. Sev, severe. Mod, moderate. Filled cells indicate that the characteristic is present in the patient, with color indicating the phenotype category as in Fig. 4.4. "NA" indicates that the patient was not assessed for the trait.

4.5 Discussion

In using two models of *de novo* mutation prediction, we identified evidence for DNM burden in 253 total genes, including 145 with a burden of LGD mutations and 123 with a burden of missense mutations. This allowed us to both identify more potential NDD risk genes and compare model performance. The top two modules of the union set of genes, identified with MAGI based on PPI and coexpression, are enriched for transcription regulation (Module 1) and neurotransmission (Module 2), confirming findings from previous studies[25,28,39,49,65,75,135–137,143,144]. Interestingly, genes with LGD burden are also enriched for transcription regulation, and genes with missense burden are enriched for neuronal function. This may indicate that mutation type is a factor in module identity, and that the differing types of mutation have different roles in disease pathogenesis.

Although there was consistency between the two models for genes with LGD burden, there was less agreement for genes with missense burden. This may be due to the variable effect of missense mutations. Whereas LGD mutations are expected to have the same functional consequence, namely loss of function, missense mutations can have effects ranging from none to loss or gain of function. Missense mutations can also have a dominant negative effect, as seen in a recurrent mutation in glutamate receptor subunit gene *GRIA1*[126]. The impact of a missense mutation is determined by the specific substitution and the location of the mutation. As the two models used in this integrated analysis calculate a relative mutation rate across the entire gene, they lose the ability to identify sub-genic regions of burden, a known disease-associated phenomenon such as the clustering of *de novo* missense mutations in Schinzel-Giedion syndrome[99].

Given the limitations in the whole-gene burden analysis for missense mutations and the potential importance of clustered missense mutations in NDDs, we used a different method, CLUMP[51], an unsupervised clustering algorithm. We find that 183 genes have *de novo* missense mutations that are significantly more clustered than controls. Interestingly, although this set of genes is largely distinct from the 123 genes detected with recurrent missense burden, function and expression enrichments of the two sets are similar. Notably, both sets of genes are enriched for expression in the cortex during fetal development. This is consistent with previous findings that implicate midfetal cortical projection neurons in ASD pathogenesis[145]. Commonalities in enrichments of the two distinct gene sets suggest that the two metrics may be converging on a network of related genes.

Despite the limitations of burden models for missense mutations, the increasing size of denovo-db has allowed identification of novel genes. Previously, the six *de novo* missense mutations observed in *TRRAP* in denovo-db (v.0.9) were insufficient to establish burden (**Table 2.1**). In denovo-db v.1.5, the number of missense mutations has doubled and the gene is now found by both models to have a significant burden of *de novo* missense events. This gene is further identified by CLUMP as having significant clustering relative to controls ($q = 1.99 \times 10^{-4}$). Of note, only one *de novo* LGD mutation has been identified in this

gene to date. Without focusing on missense mutations with increasing sample sizes, this gene may have been missed.

Although *TRRAP* has not yet been implicated in NDDs, its function as a component of chromatin complexes with histone acetyltransferases has long been associated with disease[146–148]. In addition to the significant burden and clustering of missense mutations in this gene, two individuals have a substitution at the same amino acid, and subsequent targeted sequencing of this residue identified novel variants in the surrounding area. In collaborating with an international group of researchers and clinicians, we showed a strong genotype–phenotype correlation, where patients with mutations in one of two clusters have a more severe phenotype. This illustrates the importance of assessing qualities of missense mutations that determine their functional effect, such as location. For this gene and likely others, specific missense mutations are critical in determining disease. Additionally, the significant burden of substitutions in Cluster 1 ($p_{adj} = 5.52x10^{-14}$) and the resulting severe phenotype suggest that this region of the protein, although uncharacterized, has a critical role and warrants functional follow-up.

Using both overall gene burden and clustering analysis to identify genes, we are able to identify novel NDD risk genes. As illustrated with *TRRAP*, clustering of missense mutations in some genes may be indicative of phenotypic outcome. Before investing in laborious and expensive functional experiments, more evidence for the genetic association of a gene with disease can be ascertained with targeted sequencing. Given the relevance of clustered missense mutations, we can more efficiently acquire meaningful data by targeting regions where pathogenic mutations cluster.

Chapter 5. Novel and known pathogenic missense mutations identified by targeting clusters of *de novo* missense mutation.

Data in this chapter are being prepared for publication.

5.1 Summary

The majority of genes associated with sporadic neurodevelopmental disorders (NDDs) involve *de novo* loss-of-function mutations, although it is predicted that *de novo* missense mutations will contribute as much, if not more, to the genetic etiology of NDDs. As most missense mutations are incidental, new signatures of disease risk are needed to discover pathogenic events. We used an unsupervised clustering algorithm on 11,459 cases with NDDs to identify 220 genes with *de novo* missense mutations that are more clustered than controls. This set of genes contains 63 recurrent sites and is enriched for chromatin modification and aspects of neuronal function. We targeted these regions, other clustered regions, and additional recurrent sites in 16,380 individuals with idiopathic NDDs and identified additional events at 20 targeted sites and 28 residues with only one mutation in our discovery dataset. Eighteen of these sites are known to be pathogenic and 30 are novel, although some have evidence for disease risk. Further assessment of these novel sites will provide increased understanding of the genetic architecture of NDDs.

5.2 Introduction

With the availability of sequence data from large NDD cohorts, over 200 potential risk genes have been identified by assessing gene-level *de novo* burden[149] (**Figure 4.1**). While there is convergence between two statistical models on genes that contribute to disease via LGD mutations, there is a lack of agreement for genes identified via missense burden. The limited understanding of genes with missense burden is further elaborated by looking at predictions of the number of genes implicated in NDDs. The number of genes identified with a significant burden of LGD mutations or highly deleterious (CADD[90] score > 30) missense mutations is reaching a plateau, and few additional genes are expected to be identified with

increasing cohort sizes. However, for missense mutations without a deleteriousness cutoff, the expected number of genes cannot be projected[149]. Further, while missense mutations account for more cases of ASD than LGD mutations[28], more genes with LGD burden have been identified (145 vs. 123).

The contribution of missense mutations to disease is particularly interesting because of the additional mechanisms by which these mutations can affect protein function. They can be so damaging that they cause loss of function, and they can also have gain-of-function or dominant negative effects. They are therefore potential targets for treatment and understanding of disease pathogenesis. Unfortunately, most missense mutations are incidental[28].

To better understand the genetic architecture of missense risk in NDDs, methods that distinguish deleterious from incidental mutations are needed. We previously used amino acid recurrence, a phenomenon seen in disease[92], to identify hotspots of missense mutations[126]. By refining our statistical model to focus on individual codons, we found significant burden for *de novo* missense mutations in seven codons. Targeted sequencing of these sites enabled discovery of additional recurrent events, including known pathogenic events and novel events predicted to be pathogenic. Further, functional and clinical studies of a novel recurrent event in glutamate receptor *GRIA1* that was identified in additional patients with targeted sequencing support the role of this site mutation in disease (Chapter 3). This gene is likely to have gone unnoticed without this focus on recurrent sites, as few mutations have been identified in this gene. With targeted sequencing and development of a model to test burden for individual codons, we found sufficient evidence to attribute risk to this novel gene.

Another phenomenon seen in disease, such as Schinzel-Giedion syndrome[99], and in denovo-db[79] is mutation clustering. Further support for using this feature to identify novel risk genes comes from the gene *TRRAP*, in which patient phenotypes segregate in a similar pattern to *de novo* missense mutation clustering[140]. Additionally, four recurrent sites are present within the clusters in this gene. With CLUMP, we were able to identify 182 additional genes that have significant mutation clustering in patients with ASD and DD/ID (Chapter 4). In this chapter, we include patients with EPI to identify more potential risk

genes. As functional studies can be resource-prohibitive, we then used targeted sequencing in 16,380 individuals with idiopathic NDDs to find additional mutations in candidate genes. New findings will aid in distinguishing risk genes for further follow-up.

5.3 Methods

denovo-db. In addition to the 5,624 ASD cases and 5,303 DD/ID cases studied in Chapter 4 (**Table 4.1**), we included 532 EPI cases from five cohorts[57,68,69,73,84], bringing the total sample size to 11,459 cases (**Table 5.1**). Variants were annotated on RefSeq transcripts in GRCh37/hg19 with SeattleSeq[82] version 138. *De novo* events shared by identical twins in Michaelson et al. (2012) were considered to be a single event. As described previously (Chapter 2), if multiple transcripts had missense mutations, we chose the transcript with the most missense mutations in cases and controls. We excluded mutations seen at high frequency (MAF > 0.1%) in two control populations: 1) NHLBI GO ESP Exome Variant Server (Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP), Seattle, WA (http://evs.gs.washington.edu/EVS/) [August 2017]) (N = 6,503) and 2) ExAC[93] v.0.3 without neuropsychiatric disorders (N = 45,376). Across these NDDs, 9,712 *de novo* missense events were annotated for analysis (**Table 5.1**).

Table 5.1. denovo-db v.1.5 with epilepsy cases.

Study diagnosis	Cohorts	Cases	*De novo* missense mutations*	*De novo* LGD mutations
ASD	8	5,624	4,297	911
DD/ID	5	5,303	5,149	1,446
EPI	5	532	266	76
TOTAL	18	11,459	9,712	2,433

*Missense mutations with minor allele frequency (MAF) < 0.1% in ESP (N = 6,503) and ExAC (N = 45,376)

Missense clustering and statistical analyses. denovo-db[79] v.1.5 was analyzed with CLUMP[51] (CLUstering by Mutation Position; https://github.com/karchinlab/clump) to assess mutation clustering. We

compared mutations in individuals with NDDs (N = 11,459; 9,712 *de novo* missense mutations) with two control datasets: 1) missense mutations (MAF < 1%) from Europeans (N = 420; 196,260 mutations) from 1KG[104] and 2) private missense mutations present in individuals from ExAC[93] v.0.3 without neuropsychiatric disorders (N = 45,376; 1,466,439 mutations). All variants were re-annotated using the CRAVAT software to enable exact transcript comparisons[105].

To identify genes and regions with significant burden of *de novo* missense mutations, we used the CH model[39] and assumed a mutation rate of 1.8 *de novo* coding variants per generation[85]. As outlined previously (Chapter 2), we applied the same method to individual codons to test for burden at sites. For whole genes, we utilized the Benjamini-Hochberg FDR correction based on 19,008 genes in the human genome[86] and report q-values. For sites, we used the Bonferroni FWER correction based on the number of amino acids in the genome (N = 1.1×10^7) and report the adjusted p-value (p_{adj}) for genome-wide significance. Gene ontology enrichment was assessed using PANTHER GO slim biological process annotation (database 2019-01-01) with two-tailed binomial tests and Bonferroni FWER correction, reported as p_{adj}. Expression enrichment was tested with CSEA[134] and TSEA[134] with one-tailed Fisher's exact tests. Benjamini-Hochberg corrected p-values are reported as q-values.

Targeted sequencing. Single-molecule molecular inversion probes (smMIPs[43]) were designed with the MIPgen program[97] to target clusters and sites of interest. We achieved successful capture for 205 of the 220 genes with significant clustering in denovo-db[79] v.1.5 (**Table 5.2**). Additionally, we targeted 55 genes that showed significant clustering in earlier versions of denovo-db. smMIPs for all clusters covered the exon or exons that contained the clustered mutations. We also wanted to capture 84 recurrent sites in 71 genes. Sixty-four of these sites (in 53 genes) were part of a targeted cluster, and we designed additional smMIPs for the remaining 20 sites (in 18 genes) (**Table 5.2**). For both clusters and sites, we used smMIPs targeting both strands to maximize coverage. We captured 174 kilobases of sequence in a total of 278 genes with 2,216 smMIPs.

Table 5.2. Genes with clusters and sites targeted with smMIPs.

		Genes without recurrent site	N	Genes with recurrent site	N
Significant clustering ($p < 0.05$)	**denovo-db v.1.1-v.1.3**	*ACAD10, ATP1A3, C14orf93, CHGA, CIDEC, CLCN4, CMIP, CNTNAP4, DAK, DIDO1, DNAH17, FBN1, FZD10, GLRA3, GRIN1, HIBCH, HSD3B1, HSP90AA1, INSRR, KRT27, KRTAP13-3, MAP3K7, MAPKBP1, MAST3, MYO18A, NUP188, OAZ2, P2RX1, PEPD, PRKDC, PSME4, RAB11B, RAD18, RFX2, RNF126, ROCK1, SETD1A, SH2B1, SLC9A3, SMC3, TBR1, TRAF7, TRIOBP, TUB, UACA, VCP, VPRBP, WBSCR17, WDFY4*	49	*ACTL6B, CDKL5, CHD3, CSNK2A1, FBN2, PIK3CA*	6
	denovo-db v.1.5 only	*ACACB, AP5Z1, ARHGEF1, CACNA1C, CACNA1H, COL5A2, CPSF3, CR1, CTR9, CYP26B1, DMXL2, DNAH2, FCGR2B, FHOD3, GABRP, GNB2, HMGCL, IGF2R, IQCE, ITGA7, KIAA1217, LINS, LONP1, METTL16, MGAT5B, MYO9A, NBEAL2, NFE2L2, NHLRC1, PACS1, PATL1, PBX1, RAI1, RAPGEF2, RBM19, RTL1, RYR3, SFTPB, SUSD4, TCERG1L, TNK2, UMODL1, UNC45B, ZNF333*	44	*CEP131, CEP350, EEF1A2, GABRG2, MED12, PRKAR1B, SCN8A, STK39, TSPYL1, XPO5*	10
	Shared	*ABCG2, ABI3BP, ABR, AGO2, AGO3, AGT, AKAP9, ALK, ANKHD1, ARID1A, ATIC, BCL11A, C17orf53, C1QTNF6, CBL, CBLL1, CD59, CEP170B, CHD7, CHD9, CHRNA4, CLCA2, COL4A3BP, COL6A5, CUX2, DCDC1, DEAF1, DENND4B, DNMT3A, DPF2, DSP, FAM84A, FBXO21, FGA, FOXP1, GABRA1, GALNTL5, GON4L, GPN1, HCN3, HNRNPU, ITPR1, KAT6B, KCND3, KCNK3, KDM3A, KDM5C, KLHL14, KLHL17, KMT2C, KRBA1, LAMA1, LIPT2, LRFN2, MANBA, MAP2K6, MAST1, MEF2C, MEFV, MIDN, MLH1, MTF2, MYBBP1A, MYH14, NDRG4, NPR2, NR4A2, NSL1, PCDH10, PCNX, PFKP, PHACTR1, PHF3, PIK3R2, PLD5, PLEKHA7, POLE, PRB4, PRUNE2, PTPN11, RBL2, RBM12, RGL1, RGS12, SELO, SH3BP2, SHB, SLC35C2, SLC9A3R2, SLCO1B1, SLITRK5, SLX4, SMARCD1, SOX15, SPG11, STAG1, STIL, STIM2, STXBP5, SYT1, TBC1D20, TEKT3, TNPO2, TTLL5, TXNRD1, UBQLN3, WDR26, YWHAG, ZFP69, ZFYVE16, ZNF215, ZNF318, ZNF420, ZNF7*	114	*AFF3, AKT3, ARID1B, ATP1A2, ATP6V0A1, BTF3, CDK13, COPG1, CSNK1E, CTCF, DDX23, EPHA10, EZH2, FAM104A, GABBR2, GRIA1, GRIN2A, IGHMBP2, KCNH1, KCNQ2, MAP2K1, MECP2, OGFOD3, PACS2, POU3F3, PPP2R1A, PPP2R5D, PRKD1, RAB11A, SCN3A, SHOC2, SIN3B, SMAD4, STXBP1, TMEM26, UBR2, UBTF*	37
	No clustering			*CNOT3, DDX3X, DNAH3, DPYSL5, FBXO11, GNAI1, HECW2, HIF3A, IRS2, KIF1A, NLGN2, PAX3, PHIP, PTEN, PURA, SCN2A, SIX3, TCF4*	18
		TOTAL	**207**		**71**

N, number of genes in each category.

Probes were used on 16,830 cases with NDDs, including 7,261 from ASD cohorts, 3,560 from DD cohorts, and 6,009 from cohorts that included both diagnoses, as well as 909 unaffected controls (**Table 5.3** and **Appendix 5.1**). Reads were aligned to GRCh37/hg19 with BWA-MEM[100]. Rare missense variants are those that have been seen zero or one times in gnomAD[93] exomes (N = 104,068) and genomes (N = 10,636). Only nucleotides where at least 90% of gnomAD samples had at least 10X coverage were included. Missense variants of interest in cases and controls were validated with Sanger sequencing. All patients were consented for sequencing and recontacting for inheritance testing at the providing

laboratory. Patient samples were acquired from Adelaide (Jozef Gecz, University of Adelaide), Antwerp (Frank Kooy, University of Antwerp), Autism Clinical and Genetic Resources in China (ACGC; Kun Xia), the Autism Genetic Resource Exchange (AGRE), Autism Phenome Project (APP; David Amaral, University of California, Davis), Iowa (Jacob Michaelson, University of Iowa), Leiden (Gijs Santen, Leiden University Medical Center), Leuven (Hilde Peeters, University Hospitals Leuven), Murdoch (Ingrid Scheffer, University of Melbourne), Philadelphia (Hakon Hakonarson, Children's Hospital of Philadelphia), Prague (Zdeněk Sedláček, University Hospital Motol), San Diego (Eric Courchesne, University of California, San Diego), Simons Simplex Collection (SSC), Stockholm (Magnus Nordenskjold, Karolinska University Hospital), the Study of Autism Genetics Exploration (SAGE; Raphe Bernier, University of Washington), The Autism Simplex Collection (TASC), and Troina (Corrado Romano, Associazione Oasi Maria Santissima).

Table 5.3. Targeted sequencing cohorts.

Cohort (Principal Investigator)	Cases	Primary diagnosis	Controls
Autism Clinical and Genetic Resources in China (ACGC) (Xia, Kun)	672	ASD	
Autism Genetic Resource Exchange (AGRE)	1661	ASD	97
Autism Phenome Project (APP) (Amaral, David)	145	ASD	53
Iowa (Michaelson, Jacob)	195	ASD	277
Leuven1 (Peeters, Hilde)	904	ASD	
Leuven2 (Peeters, Hilde)	988	ASD	
Murdoch2 (Scheffer, Ingrid)	56	ASD	
Philadelphia (Hakonarson, Hakon)	1445	ASD	
Study of Autism Genetics Exploration (SAGE) (Bernier, Raphe)	458	ASD	61
The Autism Simplex Collection (TASC)	737	ASD	
Antwerp (Kooy, Frank)	900	ASD/DD	
Leiden (Santen, Gijs)	210	ASD/DD	
Prague (Sedláček, Zdeněk)	384	ASD/DD	
San Diego (Courchesne, Eric)	567	ASD/DD	325
Stockholm (Nordenskjold, Magnus)	1499	ASD/DD	
Adelaide1 (Gecz, Jozef)	1246	DD	
Adelaide2 (Gecz, Jozef)	960	DD	
Adelaide3 (Gecz, Jozef)	1440	DD	
Adelaide4 (Gecz, Jozef)	839	DD	
Troina1 (Romano, Corrado)	798	DD	
Troina2 (Romano, Corrado)	285	DD	
Troina3.2016 (Romano, Corrado)	192	DD	
Troina3.2017 (Romano, Corrado)	161	DD	
Troina4 (Romano, Corrado)	88	DD	
Simons Simplex Collection (SSC)	0	NA	96
TOTAL	**16830**		**909**

5.4 Results

Clustering of *de novo* missense mutations in denovo-db. For a gene to be analyzed by CLUMP[51], it must contain at least two mutations in cases and two in controls. Of the 1,916 genes with two or more mutations in cases in denovo-db v.1.5, 1,875 and 1,789 genes contained two or more mutations in ExAC[93] and 1KG[104] control datasets, respectively. Comparison with ExAC and 1KG identifies 220 genes with mutations that are significantly more clustered than controls ($p < 0.05$, CLUMP, one-tailed permutation test), 94 of which are significant in both comparisons (**Appendix 5.2**). Sixty-six of these genes were newly identified with the increased sample size of denovo-db v.1.5, while the remaining 154 were previously identified in denovo-db v.1.2 (**Appendix 3.8**). Additionally, 72 new genes reached significance with the inclusion of cases with EPI (**Appendix 5.2**). Notably, no LGD mutations are observed in 172 of the 220 genes (78.2%) with significant clustering. Further, 31 genes with significant clustering carry a significant burden of *de novo* missense mutations (**Table 5.4**).

Similar to genes that have been associated with NDDs via burden of missense mutations (Chapter 4.4), this set of 200 clustered genes is enriched for functions related to membrane potential regulation (12 observed vs. 1.5 expected, 7.99-fold enrichment, $p_{adj} = 8.07 \times 10^{-5}$) and not transcription regulation. The expression pattern of these genes during brain development also reiterates that seen in genes with a burden of missense mutations (**Figure 4.3b**), with added enrichment for fetal striatal expression (**Figure 5.1a**). Additionally, it is enriched for expression in striatal medium spiny neurons that express dopaminergic D1 ($p_{adj} = 0.002$) and D2 ($p_{adj} = 0.002$) receptors (**Figure 5.1b**). The increased striatal expression may be driven by the additional genes identified with the inclusion of EPI in this chapter. Of the 19 genes expressed in D1 neurons, nine were not found in Chapter 4 without EPI cases (*CACNA1C, CACNA1H, FAM84A, FOXG1, PIK3R2, BCR, FOXP1, SHB,* and *ITPR1*), and for the 18 genes expressed in D1 neurons, eight were not found (*AKAP9, RAI1, CACNA1H, BCR, FOXG1, SHB, FOXP1,* and *ITPR1*).

Table 5.4. Genes with significant clustering and burden of *de novo* missense mutations.

Gene	Protein ID	Case *de novo* MIS (N = 11459)	Case *de novo* LGD (N = 11459)	Case CLUMP score	ExAC v0.3 CLUMP score difference (N = 45376)	NDD vs ExAC p-value	1KG CLUMP score difference (N = 420)	NDD vs 1KG p-value	*De novo* MIS p-value	Benjamini-Hochberg corrected p-value*
PPP2R5D†	NP_006236.1	17	2	1.100	2.818	0.001	0.217	0.037	1.73E-34	3.29E-30
STXBP1†	NP_003156.1	15	7	1.523	2.320	0.02	0.542	0.219	1.42E-17	8.98E-14
TBL1XR1	NP_078941.2	8	5	1.396	1.756	0.026	0.062	0.348	1.11E-15	4.22E-12
KCNQ2†	NP_742105.1	23	2	0.396	3.692	$<1x10^{-3}$	3.535	$<1x10^{-3}$	7.09E-14	1.68E-10
SCN8A†	NP_055006.1	14	1	1.053	3.165	0.027	1.162	0.197	2.92E-12	4.62E-09
CTCF†	NP_006556.1	8	1	0.898	2.826	0.003	1.637	0.001	6.17E-12	8.37E-09
PACS1	NP_060496.2	11	0	0	3.787	$<1x10^{-3}$	1.771	0.016	7.26E-12	9.20E-09
ANKHD1-EIF4EBP3	NP_065741.3	3	0	0.462	5.023	0.006	2.514	0.369	1.74E-12	3.31E-08
GRIN2B	NP_000825.2	14	2	1.338	3.392	0.004	1.410	0.15	5.04E-11	5.99E-08
MAP2K1†	NP_002746.1	6	0	0.324	2.952	$<1x10^{-3}$	3.952	0.003	1.19E-10	1.19E-07
COL4A3BP†	NP_001123577.1	8	0	0.771	0.968	0.348	2.354	$<1x10^{-3}$	2.01E-10	1.82E-07
MECP2†	NP_001104262.1	9	11	0.370	3.207	$<1x10^{-3}$	N/A	N/A	2.91E-10	2.51E-07
CDK13†	NP_003709.3	12	2	1.719	2.993	0.048	0.054	0.509	6.92E-10	5.06E-07
SNX5	NP_689413.1	2	0	2.565	0.896	0.047	-0.530	0.187	7.36E-10	5.18E-07
SMARCD1	NP_003067.3	2	1	1.609	1.596	0.018	0.870	0.032	1.93E-09	1.22E-06
WDR26	NP_079436.4	2	2	1.946	1.397	0.032	-1.118	0.072	2.61E-09	1.60E-06
HNRNPU	NP_114032.2	2	7	2.251	1.178	0.061	1.038	0.027	5.04E-09	2.82E-06
FOXG1	NP_005240.3	11	3	0.563	1.759	0.055	0.998	$<1x10^{-3}$	9.45E-08	4.49E-05
PTPN11†	NP_002825.3	7	0	0.813	2.314	0.03	1.686	0.003	1.03E-07	4.65E-05
MEF2C	NP_001180276.1	6	4	0.881	1.569	0.071	3.555	0.023	1.40E-07	6.19E-05
DNMT3A	NP_783328.1	9	2	1.401	2.836	0.014	1.762	0.044	6.95E-07	2.81E-04
SMAD4†	NP_005350.1	5	0	0	2.160	0.04	5.382	$<1x10^{-3}$	6.05E-06	1.95E-03
BTF3†	NP_001032726.1	4	0	0	1.633	0.024	1.891	0.001	1.92E-05	4.34E-03
CBL	NP_005179.2	5	1	1.337	3.019	0.035	0.490	0.519	7.23E-05	8.23E-03
YWHAG	NP_036611.2	4	0	0.520	1.250	0.148	1.270	0.001	8.00E-05	8.83E-03
SYT1	NP_001129278.1	4	0	1.180	2.405	0.014	0.818	0.045	1.32E-04	1.40E-02
PPP2R1A†	NP_055040.2	4	0	0.173	1.532	0.252	2.721	$<1x10^{-3}$	1.50E-04	1.56E-02
MTF2	NP_031384.1	3	0	0.799	2.715	0.116	4.456	0.034	2.21E-04	2.20E-02
PBX1	NP_002576.1	4	1	1.665	1.145	0.306	3.550	0.047	3.54E-04	3.41E-02
KCND3	NP_004971.2	4	0	1.171	2.565	0.024	1.450	0.045	3.71E-04	3.54E-02
RAB11A†	NP_004654.1	3	0	0	2.216	0.01	N/A	N/A	4.07E-04	3.83E-02
NR4A2	NP_006177.1	4	0	1.125	2.718	0.011	-0.126	0.15	5.42E-04	5.00E-02

*Benjamini-Hochberg FDR correction for 19,008 genes

†Gene has a recurrent site (two or more *de novo* missense mutations at a codon in unrelated cases in denovo-db v.1.5) not seen in ExAC

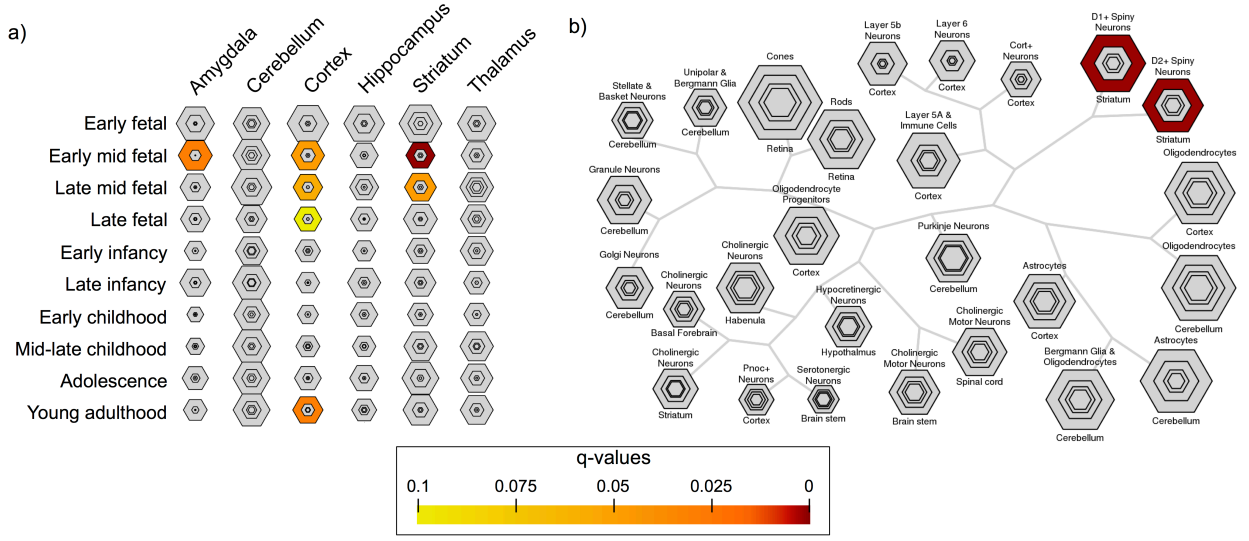

Figure 5.1. Gene expression enrichments for genes with clustered de novo missense mutations. a) Enrichments in brain regions during development[134]. Similar to genes with a burden of *de novo* missense mutations (**Figure 4.3b**), genes with clustered missense mutations in patients with ASD, ID/DD and EPI are enriched in amygdala and cortex during fetal development. This expanded set of genes is also enriched for expression in the striatum during fetal development. **b)** Enrichments in neuronal subtypes. Clustered genes are enriched for expression in medium spiny neurons in the striatum that express dopaminergic D1 and D2 receptors. This same enrichment is present in genes with a burden of missense mutation (Chapter 4.4), albeit at a much lower level. The addition of EPI cases may drive this enhanced striatal signal.

Interestingly, 50 genes with significant clustering have a recurrent site, i.e., *de novo* missense mutations affecting the same amino acid in two or more unrelated cases, some of which are known to be pathogenic. However, due to the locus heterogeneity of NDDs, only eight of the sites carry a significant burden of *de novo* missense mutations. With targeted sequencing of additional cohorts, we expect to find additional variants that will provide evidence to the association of specific recurrent mutations with NDDs. Further, by targeting all clusters, we expect to find novel recurrent mutations in these hotspots of missense mutation, especially as they overlap with known functional domains in proteins with relevant neurobiological function.

Targeted sequencing of sites and clusters. With 2,216 smMIP probes, we were able to capture 174 kilobases of sequence across clustered regions in 260 genes, 53 of which included a recurrent site, and an additional 18 recurrent sites in genes that did not have significant clustering (**Table 5.2**). We used these probes on 16,830 individuals with idiopathic NDDs, primarily ASD and DD, as well as 909

unaffected controls (**Table 5.3**). In this targeted sequencing experiment, we identified 22,747 missense variants in cases and 1,103 in controls. Filtering for rare (AC ≤ 1 in gnomAD controls (N = 114,704) at nucleotides where ≥90% of controls have ≥10X coverage) events that are predicted to be deleterious (CADD ≥ 15) leaves 1,792 missense variants in cases and 52 in controls. The variants identified can be classified based on their occurrences in denovo-db v.1.5. Of most interest are Type 1 and 2 events, especially those that have not been seen in the large gnomAD control population. Type 1 events are already established recurrent sites for which targeted sequencing has identified additional events in distinct individuals. Type 2 events are present once in denovo-db v.1.5 and were identified one or more additional times with targeted sequencing, making them potentially novel recurrent sites. All Type 1 and 2 events were validated with Sanger sequencing. Type 3 events are those that are seen two or more times with targeted sequencing but not in denovo-db. As the inheritance of these events is unknown, and they have not yet been seen *de novo* in other cases, they are a lower priority. However, the multiple observations may make them candidates for later follow-up. Finally, Type 4 events are those that have been identified singly with targeted sequencing. These events may be interesting when they have other markers of pathogenicity, such as high CADD scores[90] and location within clusters or known functional domains.

We identified 25 new Type 1 events at 20 denovo-db sites and 36 Type 2 events at 28 residues. Although inheritance information is not available for the new events, none of these missense mutations have been seen in controls but have been seen *de novo* at least once in cases. Eighteen of these mutations are known to be pathogenic for NDDs (**Table 5.5**). For the remaining 30 mutations, pathogenicity has not been established but evidence has begun to mount that associates them with disease. Eleven of the genes with novel events (**Table 5.6**) have been associated with NDDs via germline missense mutation. *ATIC* has also been determined to be pathogenic for NDDs, although the individual is a compound heterozygote at different residues[150]. Mutations in *RBM12* and *ALK* have been attributed to susceptibility for brain-derived disease[151–155], indicating that they have roles in brain function.

Table 5.5. Pathogenic events identified with smMIPs and validated with Sanger sequencing.

Site	Chr	Pos	Ref	Alt	Event	denovo-db v.1.5 (N=11,459)	smMIPs (N=16,830)	Cohort of newly identified event (primary diagnosis)	Known pathogenicity of germline mutation (inheritance)
MECP2_133	X	153296882	G	A	R133C	4	2	Troina1 (DD), Troina3.2016 (DD)	Rett syndrome (XLD)[156]
CSNK2A1_198	20	472926	T	C	K198R	4	1	Antwerp (ASD/DD)	Okur-Chung (AD)[157]
STXBP1_190	9	130425622	C	T	R190W	3	2	Prague (ASD/DD), Troina3.2016 (DD)	EIEE4 (AD)[72,158]
KIF1A_307	2	241715306	C	T	R307Q	3	1	Troina1 (DD)	MR AD 9[159]
SHOC2_2	10	112724120	A	G	S2G	2	3	Antwerp (ASD/DD), Adelaide1 (DD), Troina3.2016 (DD)	Noonan-like syndrome with loose anagen hair (AD)[160]
CDK13_717	7	40039066	G	A	G717R	2	1	Troina3.2016 (DD)	Syndromic NDD (congenital heart defects, dysmorphic face features, ID; AD)[161,162]
PTEN_246	10	89717712	C	T	P246L	2	1	TASC (ASD)	PTEN hamartoma tumor syndrome, Cowden syndrome 1 (AD)[163]
SCN2A_1773	2	166245634	C	T	A1773V	2	1	SAGE (ASD)	NDD with ASD, ID, EPI (AD)[164]
STXBP1_292	9	130430439	G	A	R292H	2	1	Stockholm (ASD/DD)	EIEE4 (AD)[165]
TCF4_576	18	52896219	G	A	R576W	2	1	Troina1 (DD)	Pitt-Hopkins syndrome (AD)[166]
PPP2R5D_200	6	42975009	G	A	E200K	1	2	Adelaide3 (DD), Adelaide4 (DD)	MR AD 35[167]
MAP2K1_124	15	66729163	C	T	P124L	2	1	Leuven (ASD)	Cardiofaciocutaneous syndrome (AD)[168]
CLCN4_718	X	10188877	C	T	R718W	1	1	TASC (ASD)	CLCN4-related disorder (XLD)[169]
EEF1A2_124	20	62126409	C	T	E124K	1	1	Troina1 (DD)	EIEE33 (AD)[170]
KCNH1_496	1	210977485	C	T	G496R	1	1	Antwerp (ASD/DD)	Zimmermann-Laband syndrome 1 (AD)[171]
KCNQ2_563	20	62044879	C	T	D563N	1	1	Antwerp (ASD/DD)	EIEE7 (AD)[172]
MAP2K1_130	15	66729181	A	G	Y130C	1	1	Troina1 (DD)	Cardiofaciocutaneous syndrome (AD)[173]
MECP2_306	X	153296363	G	A	R306C	1	1	Adelaide2 (DD)	Rett syndrome (XLD)[174,175]

AD, autosomal dominant. AR, autosomal recessive. XLD, X-linked dominant. XLR, X-linked recessive.
EIEE, early infantile epileptic encephalopathy. MR, mental retardation.

Table 5.6. Novel events identified with targeted sequencing and validated with Sanger sequencing.

Site	Chr	Pos	Ref	Alt	Event	denovo-db v.1.5 (N=11,459)	smMIPs (N=16,830)	Cohort of newly identified event (primary diagnosis)	Disorder associated with gene	Type of mutation	Inherit-ance
KCNQ2_144	20	62076675	G	A	R144W	4	1	Stockholm (ASD/DD)	EIEE7[176]	MIS	AD
KCNH1_357	1	211093375	G	A	R357W	4	1	Leuven (ASD)	Zimmerman-Laband[171], Temple-Baraitser[177]	MIS	AD
STXBP1_190	9	130425623	G	A	R190Q	3	2	Prague (ASD/DD), Troina3.2016 (DD)	EIEE4[178]	MIS + LGD	AD
RAB11A_154	15	66172039	C	T	S154L	2	1	Adelaide1 (DD)	Developmental and epileptic encephalopathy[179]	MIS	AD
GRIN2A_653	16	9923329	A	G	M653T	2	1	Adelaide2 (DD)	Epilepsy with speech disorder +/- MR[20]	MIS + LGD	AD
CTCF_368	16	67654615	C	T	R368C	2	1	Prague (ASD/DD)	MR AD 21[180]	MIS + LGD	AD
CHD7_1054	8	61735264	C	T	R1054W	1	1	Adelaide3 (DD)	CHARGE syndrome[181]	MIS + LGD	AD
TTLL5_536	14	76231013	A	G	K536E	1	1	AGRE (ASD)	Cone-rod dystrophy[182]	MIS + LGD	AR
MEF2C_33	5	88100576	A	T	Y33N	1	1	Prague (ASD/DD)	MR AD 20[183]	MIS + LGD	AD
LINS_532	15	101110121	T	A	Q532H	1	1	Troina1 (DD)	MR AR 27[184]	MIS + LGD	AR
CDK13_874	7	40102444	G	T	V874L	1	1	Stockholm (ASD/DD)	Syndromic NDD (congenital heart defects, face dysmorphia, ID)[162]	MIS	AD
ATIC_451	2	216211513	G	A	R451H	1	1	Leiden (ASD/DD)	NDD (ID, EPI, dysmorphia)[150]	compound het	AR
RBM12_6	20	34243229	G	C	R6G	2	1	Adelaide3 (DD)	Schizophrenia (susceptibility)[151]	LGD	AD
ALK_1181	2	29443675	C	T	R1181H	1	1	San Diego (ASD/DD)	Neuroblastoma (susceptibility)[155]	MIS	
DPYSL5_41	2	27121488	G	A	E41K	2	1	Troina2 (DD)	-		
UBR2_1206	6	42631075	G	A	G1206R	2	1	AGRE (ASD)	-		
ACACB_1002	12	109644607	C	A	N1002K	1	3	AGRE (ASD), Troina1 x2 (DD)	-		
MIDN_405	19	1257078	G	A	R405H	1	2	Leuven (ASD), Stockholm (ASD/DD)	-		
GLRA3_15	4	175749919	T	C	Y15C	1	2	AGRE x2 (ASD)	-		
ROCK1_753	18	18566957	T	C	Q753R	1	2	AGRE x2 (ASD)	-		
ZNF7_661	8	146068473	C	A	P661T	1	2	Adelaide2 (DD), Adelaide3 (DD)	-		
ZNF7_661	8	146068474	C	G	P661R	1	2	Adelaide2 (DD), Adelaide3 (DD)	-		
CACNA1H_308	16	1250375	G	A	R308H	1	1	Adelaide3 (DD)	-		
CSNK1E_136	22	38696888	T	A	M136L	1	1	TASC (ASD)	-		
DCDC5_536	11	30926530	A	G	I536T	1	1	CHOP (ASD)	-		
METTL16_155	17	2376824	T	C	I155V	1	1	AGRE (ASD)	-		
NBEAL2_2448	3	47049023	G	C	R2448P	1	1	Troina1 (DD)	-		
PCNX_1818	14	71555924	C	A	F1818L	1	1	AGRE (ASD)	-		
SHB_297	9	37974783	T	C	Y297C	1	1	TASC (ASD)	-		
STXBP5_425	6	147635147	C	T	R425C	1	1	Troina1 (DD)	-		

AD, autosomal dominant. AR, autosomal recessive. EIEE, early infantile epileptic encephalopathy. Het, heterozygous. MR, mental retardation.

Alternate substitutions at three of the amino acids with Type 1 events have been determined to be pathogenic for NDDs[72,185,186]. Two of these occur in potassium channels KCNQ2 and KCNH1. For both, the substitution predicted in this study is arginine to tryptophan. This substitution is nonconservative and is expected to alter secondary protein structure (GeneDx, [https://www.genedx.com/]). The two residues affected are in (KCNH1 p.Arg357) or adjacent to (KCNQ2 p.Arg144) the transmembrane domains that

provide channel function. The missense mutation in *KCNH1*, seen four times in denovo-db v.1.5 and found in one additional case with targeted sequencing, is particularly interesting as it occurs in the S4 transmembrane segment (**Figure 5.2**). This segment serves as the voltage sensor for these voltage-gated ion channels, and the substituted arginine is one of a series of positively charged amino acids that carries out this critical function[187]. While only one new mutation altering a gating residue was identified in this study, substitutions of four additional arginines that serve this function in other voltage-gated potassium channels were identified in Geisheker et al. (2017).

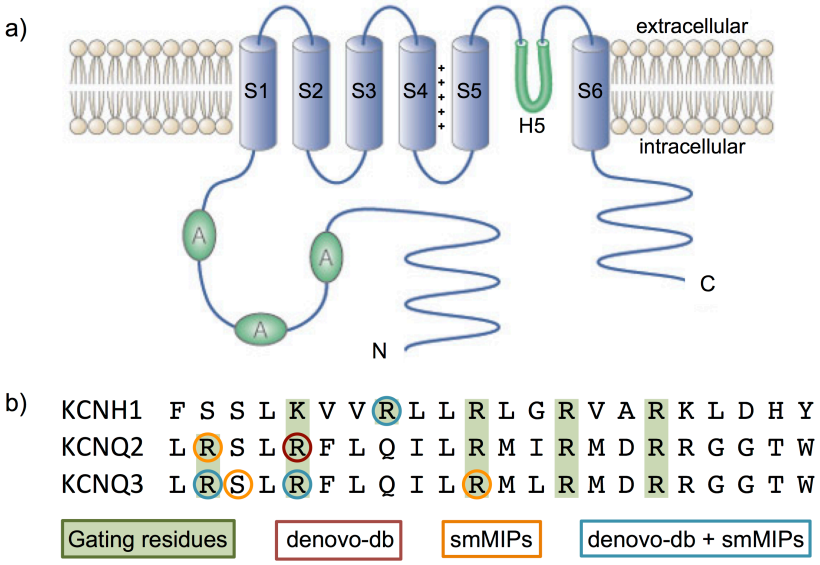

Figure 5.2. Recurrent substitutions at critical residues in voltage-gated potassium channels. a) Ligand-gated ion channels are composed of six helical transmembrane domains (S1-S6), with a reentrant pore loop (H5) connecting S5 and S6. A series of arginines and lysines in the S4 domain provides a positive charge and serves as the voltage sensor. This gating mechanism is critical to synaptic function. **b)** Missense mutations resulting in substitutions in the S4 domain of voltage-gated potassium channels in individuals with NDDs. Green boxes indicate gating residues. Red circles indicate residues substituted in denovo-db v.1.5, yellow circles indicate substitutions found with targeted sequencing with smMIPs, and blue circles indicate substitutions seen in both denovo-db v.1.5 and in new cases in our targeted sequencing cohort. The p.Arg357Trp substitution in KCNH1 was found in this study, and the events in KCNQ2 and KCNQ3 were identified in Geisheker et al. (2017).

5.5 Discussion

The goal of this study was to better elaborate the missense signal in NDDs by identifying recurrent mutations that are not seen in control populations. We used an unsupervised algorithm to find genes with clusters of missense mutation in 11,459 cases, and designed targeted sequencing probes for hotspots in 278 genes, including 71 with recurrent sites. With 2,216 smMIPs, we captured 174 kilobases of sequence

in an additional 16,380 cases with ASD and DD/ID and identified recurrent missense variants at 48 sites. Targeted sequencing again proved to be an efficient method for capturing sequence in a large number of potential candidate genes that contain hotspots.

Several lines of evidence support the relevance of these hotspots with NDDs. First, 31 genes with significant clustering have a burden of *de novo* missense mutations. Second, the set of clustered genes is enriched for relevant functions such as chromatin remodeling (12.24-fold, p_{adj} = 1.17E-03) and several aspects of brain function including regulation of neuronal membrane potential (4.82-fold, p_{adj} = 2.41E-05). Third, 63 recurrent sites are present in 50 genes with significant clustering, and many of these sites are known to be pathogenic in NDDs. This overlap between recurrent sites and clusters motivated us to use them as sequencing targets in a larger cohort.

By targeting clusters and additional recurrent sites, we found further evidence for 20 recurrent sites and established 28 new sites. Eighteen of these 48 sites are known to be pathogenic, and these new findings may help to understand the role of these mutations and genes in development. Clinically, the additional phenotypic data can provide a better understanding of developmental trajectory, which is beneficial to parents and medical providers. Assessment of common symptoms may also lead to characterization of a syndrome, which can improve diagnosis and allow earlier and more specific treatments.

The remaining 30 sites established with missense mutations in denovo-db and found with targeted sequencing are novel. Eleven of the genes with these novel sites have been associated with NDDs, indicating their role in brain development. Two of the substitutions (KCNQ2 p.Arg144Trp and STXBP1 p.Arg190Gln) have not been previously reported but, as the substitutions are nonconservative, they are predicted to be pathogenic (GeneDx, [https://www.genedx.com/]). Additionally, alternative substitutions at these two residues as well as KCNH1 p.Arg357 are considered pathogenic. As the *KCNH1* mutation alters an amino acid critical for voltage sensing and channel gating, this mutation warrants functional follow-up. Although this was the only mutation at a gating residue identified in this study, mutations at paralogous residues in other voltage-gated ion channels such as KCNQ2 and KCNQ3 have been found

with other targeted sequencing studies[126]. These events are individually rare but increased support for their role in disease is found by looking across the gene family.

As seen with potassium channels, we may be able to leverage similarities amongst other gene families to identify additional genes and mutations that contribute risk to NDDs. Multiple members of some gene families, such as zinc finger proteins and voltage-gated sodium channels, have significant clustering. And for both of these families, additional genes have *de novo* missense mutations in denovo-db v.1.5 but too few to reach significance. Assessment of burden in paralogous regions of related genes may highlight new potential risk genes. Further, known pathogenicity of mutations at specific residues, such as in potassium channels, likely warrants follow-up for novel mutations at paralogous residues in other gene family members. Given the relationship between functional importance and conservation, these sites may be particularly interesting for missense mutations. As nearly 80% of genes that are known to be associated with disease are part of gene families[188], we see this as a powerful new method to identify novel risk genes and being to understand the mechanisms of developmental neuropathology that result from such mutations.

Chapter 6. Summary and future directions

6.1 Summary of findings

In this thesis, I have shown that there is a significant signal for *de novo* missense mutations in ASD and other NDDs, and I have identified novel genes and mutations that may be causative. In Chapter 2, I aggregated mutations across several phenotypically overlapping NDDs, namely ASD, ID, DD, EPI, and SCZ. This enabled identification of genes with a significant burden of *de novo* missense mutations. I also found amino acids with multiple *de novo* substitutions in unrelated cases that were not observed in controls and developed a statistical model to test the significance of these sites. In Chapter 3, I used targeted sequencing on 17,688 individuals with idiopathic NDDs and found additional evidence for some site mutations. Functional follow-up for a mutation seen in five individuals in glutamate receptor subunit *GRIA1* showed the same changes in ion flux that cause a severe phenotype in mice. In Chapter 4, the growth of ASD and DD/ID cohorts enabled the identification of 253 genes with significant *de novo* mutation burden, including 123 with missense burden and 183 with missense mutations that are significantly more clustered than controls. With findings from targeted sequencing and an international collaboration, I showed a correlation between missense mutation clustering in *TRRAP* and patient phenotypes. Finally, in Chapter 5, I described additional recurrent site mutations found with further sequencing that targeted clustered regions. Here, I address the progress made with this research, its implications, and future directions that will advance the field of NDD genetics.

6.2. Signal for *de novo* missense mutations in NDDs

A seminal paper in 2014 on 2,517 probands with ASD and their families was the first to find a significant signal for *de novo* missense mutations[28]. While unaffected siblings had a rate of 0.82 events per child, affected children had a rate of 0.94 for an ascertainment differential of 0.12 (**Figure 1.1**). This ascertainment differential can be equated to the rate of pathogenic events in children with ASD. Importantly, it is greater than the ascertainment rate LGD mutations (0.09), meaning that missense mutations account for more cases of ASD than LGD mutations. However, due to the high rate of

incidental missense mutations, only 13% (0.12/0.94) of those seen in cases are causative. This necessitates a large sample size to observe more *de novo* missense mutations in a gene than expected. Of the 1,675 genes with *de novo* missense in this study, only 320 had two or more. This was insufficient to establish gene-specific burden. Given the utmost importance of gene identification in understanding pathogenesis, larger sample sizes are necessary to further ASD research.

In Chapter 2, we took advantage of the extensive comorbidities between the NDDs, described in Chapter 1 to achieve a dataset that was sufficiently large for burden detection. We included individuals that have CHD and a secondary NDD diagnosis, as well as those with SCZ due to known overlap in implicated genes[37]. Although the inclusion of these other disease cohorts decreased the specificity of our findings, the increased sensitivity allowed us to identify 35 genes with a burden of *de novo* missense mutations after Benjamini-Hochberg correction ($q < 0.05$), including 13 missense-specific genes (**Appendix 2.1**). As ASD and DD/ID cohorts grew, missense burden was identified in 123 genes for these diagnoses alone (**Tables 4.2** and **4.3**). While the number of genes with a burden of *de novo* LGD mutations is approaching an asymptote at 216 genes, the number of genes with missense burden cannot yet be predicted[149] (**Figure 6.1**). This suggests that the role of missense mutations in NDDs is less understood, further evidenced by the discordance between two *de novo* burden models. These findings also suggest that these models, while effective for LGD mutations, may not be as applicable for missense mutations. Further, they suggest that many more genes that contribute to NDDs via missense mutations await discovery.

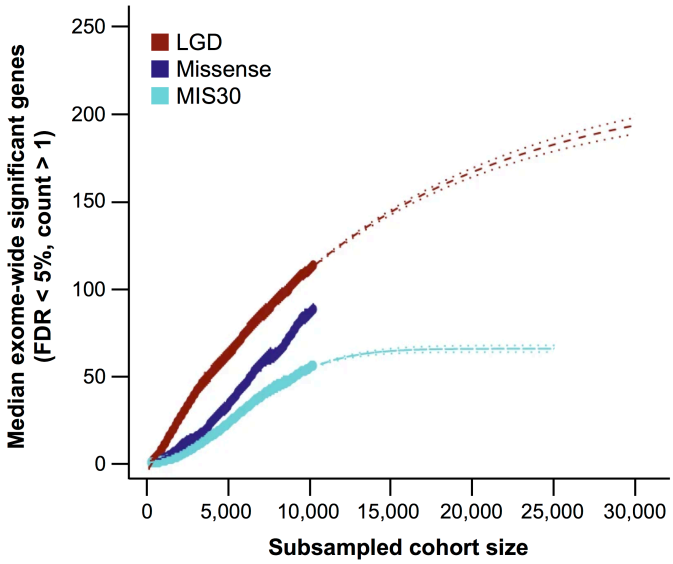

Figure 6.1. Estimation of gene discovery rates in future cohorts. From Coe et al. (2019). Using the CH model, the number of genes expected to reach significance for *de novo* mutations was estimated at increasing cohort sizes. For LGD and MIS30 mutations, the number of genes identified in 10,927 individuals in denovo-db v.1.5 with ASD or ID/DD is reaching the maximal predicted number. But for all missense mutations, including MIS30, no best-fit line was detected, suggesting that the contributions of missense mutations to NDDs are less understood.

6.3 Patterns of *de novo* missense mutation

As missense mutations are expected to explain more cases of ASD, we were motivated to find alternative methods for identifying potential risk genes. A limitation of existing *de novo* burden models is that they only assess the total number of mutations in a gene. While this is satisfactory for LGD mutations, for which functional effect is the same all along the protein, it misses a critical factor in the deleteriousness of a missense mutation — the location. Missense mutations can be highly disruptive if they fall within critical functional domains, but may have little impact elsewhere. To better account for this, we developed a modified version of the CH *de novo* burden model that predicts the number of mutations expected in a region of a protein, even down to the level of individual codons[39,126]. With this model, we found that seven codons had more *de novo* missense mutations in denovo-db (v.0.9) than expected, and one occurred in a gene that failed to reach whole-gene significance (*ALG13*). In Chapter 3, we used targeted sequencing to find 21 additional cases with missense mutations that were seen two or more times in denovo-db (**Table 3.1** and **Appendix 3.3**). Two additional codons reached significance after stringent genome-wide correction.

With targeted sequencing, we also identified mutations of interest surrounding recurrent sites (**Table 3.2**). This clustered pattern of missense mutations has been associated with NDDs, such as Schinzel-Giedion syndrome[99]. Missense mutation clustering is also a known feature of genes that cause autosomal dominant diseases[51]. We therefore used an unbiased clustering algorithm, CLUMP, to identify genes that had missense mutations that were more clustered than controls[51,126]. Preliminary results, detailed in Chapter 3, showed that this method highlighted a set of genes with relevant functional enrichments, namely aspects of neuronal communication (**Appendix 3.8**). Application of the algorithm to later versions of denovo-db (v.1.5) with increased sample sizes reiterated these findings. Interestingly, although the set of genes identified with CLUMP in Chapter 4 was largely distinct from the set of genes identified with whole-gene *de novo* missense burden, the two sets shared functional and expression enrichments. Additionally, 41 genes that have significant clustering of mutations in denovo-db (v.1.5) have been associated with NDDs (**Appendix 4.5**). Further, one of the 49 novel genes identified with this method, *TRRAP*, has a clustered mutation pattern that correlates with patient phenotypes[140] (**Figure 4.4** and **Table 4.4**). This was the first time that this clustering algorithm was used on mutations in patients with idiopathic NDDs, and these findings illustrate its utility in identifying missense-specific risk genes that are missed by whole-gene burden models.

6.4 Prioritizing genes and mutations for further investigation

The primary goal of human medical genetics research is to find mutations that cause disease so that treatments can be developed. Requirements for a mutation to be considered pathogenic are defined by the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. The only single criterion considered to be “very strong” evidence is a null variant in a gene where loss-of-function mutations are known to cause disease[189]. While this does not apply to missense mutations, several strong and moderate criteria are met by mutations in denovo-db and found with targeted sequencing in my work. *De novo* inheritance, a strong criterion, is true of all events in denovo-db, and by working with clinicians and families who provided samples for targeted sequencing, we can establish

inheritance for some newly identified events. Additionally, missense mutations identified with CLUMP meet several moderate criteria. This algorithm identified hotspots of mutation, many of which fall over known functional domains, and some of the contributing substitutions occur at residues that are known to be involved in disease. Another important source of evidence for pathogenicity is functional experiments that show the damaging effect of a mutation. Although these studies are invaluable in characterizing the role of a mutation in disease, they are often time-consuming and can be expensive.

Given the high rate of incidental missense mutations, strategies are needed to prioritize the many missense mutations identified for functional follow-up. One of the metrics that I used in my work to assess mutations for further investigation is the lack of observation of the mutation in control populations, which also serves as evidence for pathogenicity[189]. The first control population that I used, in Chapter 2, was the unaffected siblings of probands with ASD from the SSC[21]. While this dataset is a strong basis for comparison of the sibling pairs, it has limitations. First, some of the unaffected siblings have ASD characteristics, as shown by their scores on the Social Responsiveness Scale (SRS) test, which measures behaviors associated with ASD. Of the 2,032 unaffected siblings, 732 were considered concordant with their affected siblings as they had an SRS score over 50. Further, 87 of the concordant siblings had a score over 60, indicating that they have mild to moderate difficulties with social interactions[81]. We excluded siblings with a score over 60 from analyses in Chapter 2, but the classification of these individuals as controls is concerning. They may have concordant phenotypes due to inherited events, confounding analysis of *de novo* events. They also serve as a reminder that other control populations may not truly be free of disease. Secondly, the small size of this cohort limits comparisons. In Chapter 2, we used downsampling simulations to address this problem, but the recent creation of large control cohorts has provided a different avenue.

In 2014, the ExAC published exomes of 60,706 unrelated individuals[93]. The size of this dataset made it a powerful resource but the samples used came from disease cohorts. In fact, 15,330 of the individuals had neuropsychiatric diagnoses. We used a subset of the data without these individuals (N = 45,376) for comparisons in Chapters 2-5, but other individuals with NDDs could be present due to lack of diagnosis

or study exclusion. Another issue with ExAC data is the lack of information about inheritance. As a proxy, we used private (AC = 1) variants in ExAC to compare with *de novo* mutations in cases. An additional concern, especially relevant as it is a factor in determining pathogenicity[189], is the observation of known pathogenic variants, such as the mutation that causes Schuurs-Hoeijmakers syndrome[92], in this database. The mutation in this individual could be somatic or mosaic, but its presence challenges our use of control allele count as a criterion for the likelihood that a case mutation is pathogenic. Despite these limitations, this dataset and its successor, gnomAD, are excellent sources of data on human variation. gnomAD has 110,476 controls without neuropsychiatric diagnoses and includes data from whole-exome sequencing (WES) and whole-genome sequencing (WGS). All issues that were present with ExAC data are also relevant for gnomAD data, but coverage for WGS was much better than WES, enabling more complete comparisons.

A second metric that we used in evaluating mutations is the CADD score[90]. This score is a prediction on the deleteriousness of a variant. This scoring system purports to improve on limitations of singular annotation tools, such as ascertainment bias of known pathogenic variants and the need to evaluate several different methods that are not easily comparable. It was created using machine learning and was trained with information from 63 different annotation metrics. CADD scores for some known pathogenic mutations were shown to have a greater area under the curve than other metrics of mutation characterization like GerpS, PhCons, phyloP, and Grantham scores[90]. Given that these other metrics often provide disparate results, an important utility of CADD is that it presents a singular prediction score that takes into account data from the other metrics. We therefore used CADD scores as a screening factor when assessing mutations. In Chapter 2, we showed that the distribution of CADD scores for *de novo* missense mutations in cases was significantly shifted compared to controls, suggesting that case mutations are more deleterious[126]. We also used CADD scores over 30 to define a separate class of mutation in Chapter 4, as these mutations are predicted to be in the top 0.1% of deleteriousness and may therefore be so severe that they have a loss-of-function effect. In a study that determined the pathogenicity of incidental exomic findings, low CADD scores were predicted for variants determined to be benign, but variants of all pathogenicity levels (likely benign, likely pathogenic, pathogenic) had high

CADD scores[190]. Further, CADD scores as low as 12.37 were attributed to variants determined to be pathogenic[190]. As we limited our focus to events with CADD scores over 20 for follow-up in this thesis, we may therefore have missed some pathogenic events. Future studies may benefit from using CADD scores as a component of ranking variants instead of using it as a strict filtering metric.

6.5 Convergent findings implicate genes involved in synapse and transcription regulation

A consistent finding in this thesis has been the role of NDD risk genes in synaptic signaling and transcription regulation. This replicates findings throughout the field, beginning from some of the earliest studies[22,25,75,118,135,138,143,191,192]. Enrichment for both of these functions is seen in the 35 genes with a burden of *de novo* missense mutations in Chapter 2 (**Appendix 2.1**) and in genes with significant clustering in Chapters 3 (**Appendix 3.8**) and 5 (**Appendix 5.2**). The top two modules of genes, identified by MAGI[132], with any type of mutation burden in Chapter 4 also have these enrichments, the first being transcription regulation and the second being synaptic transmission (**Figure 4.2**). Interestingly, enrichments for genes with burden in denovo-db v.1.5 are similar but diverge based on mutation type – genes with LGD burden are enriched only for transcription regulation, and genes with missense burden are enriched only for synaptic function. This may indicate that mutation type is a factor in MAGI module definition, and also that the different types of mutation have different roles in disease.

In addition to genes with *de novo* missense burden, genes with significant clustering in Chapter 4 (**Appendix 4.4**), which focused just on ASD and DD/ID, are also enriched only for synaptic function. This commonality, despite the limited overlap of the two gene sets, lends further support for the role of synaptic proteins in NDDs. Additionally, both sets of genes are enriched for expression in the cortex and amygdala during fetal development (**Figure 4.3**). It is no surprise that alterations in gene expression during this critical period of brain development may lead to NDDs, but further examination is necessary to implicate specific genes and processes.

An important class of gene with *de novo* missense burden that contributes to enrichments in synaptic signaling are voltage- and ligand-gated ion channels, including potassium channels *KCNC1*, *KCND3*, *KCNH1*, *KCNJ6*, *KCNQ2*, and *KCNQ3*; sodium channel *SCN2A*; GABA receptors *GABRB2* and *GABRB3*; and glutamate receptor *GRIN2B*. These selective channels are the foundation of neuronal signaling, and disruptions in the function of several of them, such as *SCN2A* and *GRIN2B*, are known to be pathogenic[72,193]. Some (*KCNC1*, *KCND3*, *KCNH1*, *KCNJ6*, *KCNQ3*, and *GABRB2*) are missense-specific, suggesting that LGD mutations may be associated with other phenotypic outcomes or are incompatible with life. As seen in *GRIA1* (**Figure 3.2**), specific missense mutations may also have gain-of-function effects[126]. In mice with the same amino acid change in Grid2 as seen in cases in GRIA1, the substitution leads to excess ion influx, which results in excitotoxicity and cell death[113]. Similarly, missense mutations altering specific lysines and arginines in the S4 domain of potassium channels are expected to disrupt channel gating as they provide the molecular mechanism for voltage sensing[187] (**Figure 5.2**). Some of these mutations are known to be pathogenic, and by homology, others are predicted to be as well. Functional and phenotypic follow-up will further describe the potential role of these mutations in pathogenesis.

6.6 Defining subtypes of NDDs

ASD diagnosis requires only two things: deficits in social communication and restricted or repetitive behaviors with a developmental onset[1], and there are multiple exemplars of deficits within each domain. These loose criteria mean that it encompasses individuals with a range of different phenotypes. Not only do the two components vary in severity across cases, but they are often accompanied by a wide variety of comorbidities, from ID to anxiety to gastrointestinal symptoms[3]. Additionally, ASD is diagnosed based on behavior, not cause. Other NDDs, such as ID and EPI, have much more constrained diagnostic criteria but often co-occur with ASD, leading to NDD phenotypes that are highly variable across cases.

Recent analyses find that ASD has high heritability[5] (>80%) and that the genetics are highly heterogeneous, with 500 to 1000 genes predicted to be causative[27]. It has also become apparent that

different genetic events lead to different developmental trajectories and outcomes[48,126,140,194–197]. But the high rate of NDD comorbidities, along with differences in penetrance and genetic background, lead to variable phenotypes, such as with the 15q11.3 microdeletion, which is associated with ASD, ID, EPI, and SCZ amongst different patients[38]. Instead, a genotype-first approach has successfully classified subtypes of NDDs by assessing patients with shared genetic events[46]. In this thesis, I have also used this approach by identifying sites and clusters that may be associated with disease and then using targeted sequencing to increase sample size in regions of interest. Follow-up on two genes, *GRIA1* and *TRRAP*, builds support for the association between missense mutations found in these genes and NDDs and the effectiveness of this approach.

GRIA1, a subunit of ionotropic glutamate receptors, is a critical component in synaptic signaling and learning and memory formation[116]. Upon observation of two *de novo* p.Ala636Thr substitutions (NP_001244950.1) in cases in denovo-db (v.0.9), we targeted this site for sequencing in 17,600 additional cases and found the same event three more times (**Figure 3.2a**), one of which was found to be *de novo*[126]. We found that this substitution occurred in a highly conserved region of the protein known to be critical for channel gating (**Figure 3.2b**), and that the same Ala>Thr amino acid change in a different glutamate receptor subunit, *Grid2*, causes severe ataxia in mice[111]. This phenotype in mice is consistent the restricted expression of this gene in the cerebellum. With functional experiments, we showed that the *GRIA1* mutation effected the same constitutively active channel as the *Grid2* mutation does in mice[113,126]. Furthermore, detailed phenotypic information available for four of the patients showed consistencies, including ID, ASD, and specific language delays.

Correlation between genotype and phenotype was also found in missense mutations identified in *TRRAP* in denovo-db, with targeted sequencing, and through an international collaboration[140]. This gene contributes to transcription regulation through its position in histone acetyltransferase complexes[198]. It is largely missense-specific, with discovery of only one LGD mutation[79,129]. We initially targeted this gene with smMIPs[126] because two unrelated individuals had a substitution at the same highly conserved residue (p. Trp1866, NP_001231509.1). As more studies were added to denovo-db, this gene reached

significance for *de novo* missense burden and clustering (Chapter 4). The series of *de novo* missense mutations identified by collaborators, including one found with targeted sequencing, fall in a highly clustered pattern (**Figure 4.4a**). Notably, patient phenotypes segregate with mutation clustering. Individuals with substitutions in Cluster 1, from p.Ile1031 to p.Gly1159, have a severe phenotype, with ID and malformations. Individuals with substitutions in Cluster 2, from p.Arg1859 to p.Pro1932, and in non-clustered regions have a milder phenotype of ID and/or ASD. These findings further support the validity of clustering as a feature of disease risk genes and a target for future functional studies.

6.7 Future directions

Findings in this thesis have shown that *de novo* missense mutations and genes make important contributions to NDD pathogenesis. However, as evidenced by the lack of model for the number of genes with missense burden we can expect to find, more work is needed to fully understand the role of these mutations in pathogenesis. One method for discovery is the continued growth of NDD cohorts, and the Simons Foundation is accomplishing it through a project called SPARK for the Simons Foundation Powering Autism Research for Knowledge[199]. This is an ASD-specific cohort, although many cases may have comorbidities. This group aims to recruit 50,000 families with a child diagnosed with ASD. Significant progress has already been made thanks to the affordability and efficiency of sequencing.

To date, over 27,000 individuals have been sequenced, including 8,578 families. This more than doubles our ASD sample size. This preliminary dataset includes 17,493 *de novo* missense mutations, 8,623 of which occur in genes with one or more *de novo* missense mutations in denovo-db v.1.5. It also includes new mutations at four denovo-db recurrent sites, a second mutation at 50 residues with a single mutation in denovo-db, and 311 new sites with two or more mutations in SPARK. Analysis with burden models and CLUMP are likely to identify further novel risk genes. Additionally, with the increased cohort size for individuals with ASD, we may be able to identify genes that preferentially result in ASD. Although WES is now affordable enough to do at large scales and targeted sequencing may no longer be necessary, the analysis of patterns of *de novo* missense mutations is still valuable in prioritizing genes and mutations for

follow-up. WES and WGS of additional large case and control cohorts, such as gnomAD, will continue to provide valuable information on human variation and its role in disease.

An important step, especially in differentiating subtypes of idiopathic NDDs, is collection and availability of detailed phenotypic information. Many data points were available for the SSC[21], the earlier cohort created by the Simons Foundation. Unfortunately, less information is available on the SPARK cohort, and even less on others such as the Deciphering Developmental Disorders[76,91] cohort, which provides no information. Without these data, we are unable to compare across patients and identify features that may be specific to mutation pathogenesis. The use of more standardized testing and reporting in large and small cohorts will be very beneficial in determining correlations between genotype and phenotype. Not only will this help researchers, but the description and refinement of syndromic NDDs based on genetic mechanisms will help clinicians in identifying and treating patients. It will also help patients and their families, as parents can gain support and information about developmental trajectory and effective treatments from other parents.

With an increasing amount of sequence data from cases, and an increasingly large set of data from control populations for comparison, we expect that more genes will be identified with a significant burden of *de novo* missense mutations. However, some genes may still be missed due to the rarity of pathogenic *de novo* events. This is especially true with regards to the whole-gene burden models compared in Chapter 4, which are discordant in gene identification. The lack of a best-fit model for the number of genes expected with increasing cohort sizes (**Figure 6.1**) speaks to the complicated roles of missense mutations in NDDs. Instead of looking at purely quantitative measures, which are sufficient for LGD mutations, full appreciation of the role of missense mutations in NDDs requires assessments that consider qualitative features such as clustering.

In this thesis, we assessed clustering from a linear perspective, across an unfolded protein, where the only factor was protein length. However, post-translational folding can drastically alter the proximity of amino acids along the protein, and the resulting secondary and tertiary structures are critical in protein

function. Assessment of 3-dimensional clustering after folding is therefore likely to reveal novel clusters of mutation that were not visible in linear analysis with CLUMP. This was observed in PTNP11 (**Figure 3.4b**), and continued work on protein structure may enable more high-throughput assessments. Another feature of missense mutations that is important in their functional effects is their location within a protein. Interestingly, although CLUMP was an unsupervised algorithm, clusters of mutation in many genes, especially neuronal ion channels, frequently fell over known functional domains (**Figure 3.3**). Evidence from CLUMP that mutations in a specific domain are involved in disease can be leveraged to identify additional risk genes if they have mutations in similar domains. Similarly, novel mutations at residues that are paralogous to known pathogenic events in gene families[188], such as potassium channels, warrant further investigation, especially if functional experiments have demonstrated mutation effects.

Although functional experiments are unmatched in providing evidence for the role of a mutation in disease, the time and resources required often limit the scope of the work. Fortunately, technological advances in gene editing are removing some of these limitations. One recent advance is saturation mutagenesis, where a library containing all possible mutations across a region is created[200,201]. This may be useful in assessing hotspot regions that bear clusters of mutations as identified with CLUMP or other metrics, especially when the cluster falls in an uncharacterized region of a protein such as in TRRAP (**Figure 4.4a**). While high-throughput assays have been designed to study proteins with diverse functions, from kinases[202] to DNA damage repair[203], assays for synaptic signaling will be challenging. Current electrophysiological experimental techniques are the gold standard in functional mutation characterization[204] but are not scalable. An additional technique that is greatly advancing genetics research is gene editing with CRISPR[205,206]. This method is particularly useful in creating transgenic animals to study the *in vitro* effects of a mutation. While it is vastly more efficient to create animal models with CRISPR[207], it can be challenging to extrapolate phenotypic findings from animal research to NDDs, especially ASD. Certain components and comorbidities of ASD can be parallelized into animal behaviors but the core social features of ASD are innately human. Nevertheless, animal models can provide important insights into the neuronal and synaptic impact of mutations[208], and assessment of the phenotypic outcome may add to understanding of the overall mutation impact. These studies, in

conjunction with quantitative and qualitative genetic assessments in large cohorts of cases, will be important in understanding NDD pathogenesis and developing more specific therapies that can provide improved quality of life for patients and their families.

Bibliography

1. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. *Arlington* (2013). doi:10.1176/appi.books.9780890425596.744053
2. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders, 4th Ed. DSM-IV-TR*. *American Journal of Critical Care* **25**, (2000).
3. Baio, J. *et al.* Prevalence of Autism Spectrum Disorders in a Total Population Sample-Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2014. *MMWR Surveill Summ* **67**, 1–25 (2018).
4. Blomquist, H. K. *et al.* Frequency of the fragile X syndrome in infantile autism. A Swedish multicenter study. *Clin. Genet.* **27**, 113–7 (1985).
5. Sandin, S. *et al.* The heritability of autism spectrum disorder. *JAMA* **318**, 1182–1184 (2017).
6. Hallmayer, J. *et al.* Molecular analysis and test of linkage between the FMR-1 gene and infantile autism in multiplex families. *Am J Hum Genet* **55**, 951–959 (1994).
7. Klauck, S. M. Genetics of autism spectrum disorder. *Eur. J. Hum. Genet.* **14**, 714–20 (2006).
8. Kallioniemi, A. *et al.* Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science (5083).* **258**, 818–821 (1992).
9. Solinas-Toldo, S. *et al.* Matrix-based comparative genomic hybridization: Biochips to screen for genomic imbalances. *Genes Chromosom. Cancer* **20**, 399–407 (1997).
10. Koolen, D. a *et al.* A new chromosome 17q21.31 microdeletion syndrome associated with a common inversion polymorphism. *Nat. Genet.* **38**, 999–1001 (2006).
11. Sharp, A. J. *et al.* Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. *Nat. Genet.* **38**, 1038–1042 (2006).
12. Shaw-Smith, C. *et al.* Microarray based comparative genomic hybridisation (array-CGH) detects submicroscopic chromosomal deletions and duplications in patients with learning disability/mental retardation and dysmorphic features. *J. Med. Genet.* **41**, 241–248 (2004).
13. Shaw-Smith, C. *et al.* Microdeletion encompassing MAPT at chromosome 17q21.3 is associated with developmental delay and learning disability. *Nat. Genet.* **38**, 1032–1037 (2006).
14. Vissers, L. E. L. M. *et al.* Array-based comparative genomic hybridization for the genomewide

detection of submicroscopic chromosomal abnormalities. *Am. J. Hum. Genet.* **73**, 1261–1270 (2003).

15. Penrose, L. S. Parental age and mutation. *Lancet* (1955). doi:10.1016/S0140-6736(55)92305-9
16. Zoghbi, H. Y., Ledbetter, D. H., Schultz, R., Percy, A. K. & Glaze, D. G. A de novo X;3 translocation in Rett syndrome. *Am. J. Med. Genet.* **35**, 148–151 (1990).
17. Sebat, J. *et al.* Strong association of de novo copy number mutations with autism. *Science* **316**, 445–449 (2007).
18. O'Roak, B. J. *et al.* Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat. Genet.* **43**, 585–9 (2011).
19. Abrahams, B. S. & Geschwind, D. H. Advances in autism genetics: on the threshold of a new neurobiology. *Nat. Rev. Genet.* **9**, 341–55 (2008).
20. Endele, S. *et al.* Mutations in GRIN2A and GRIN2B encoding regulatory subunits of NMDA receptors cause variable neurodevelopmental phenotypes. *Nat. Genet.* **42**, 1021–6 (2010).
21. Fischbach, G. D. & Lord, C. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron* **68**, 192–5 (2010).
22. Levy, D. *et al.* Rare de novo and transmitted copy-number variation in autistic spectrum disorders. *Neuron* **70**, 886–897 (2011).
23. Sanders, S. J. *et al.* Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams Syndrome region, are strongly associated with autism. *Neuron* **70**, 863–885 (2011).
24. Iossifov, I. *et al.* De novo gene disruptions in children on the autistic spectrum. *Neuron* **74**, 285–99 (2012).
25. O'Roak, B. J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485**, 246–250 (2012).
26. Neale, B. M. *et al.* Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**, 242–5 (2012).
27. Sanders, S. J. *et al.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237–41 (2012).
28. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature*

515, 216–221 (2014).

29. Amiet, C. *et al.* Epilepsy in autism is associated with intellectual disability and gender: Evidence from a meta-analysis. *Biol. Psychiatry* **64**, 577–582 (2008).

30. Bolton, P. F. *et al.* Epilepsy in autism: Features and correlates. *Br. J. Psychiatry* (2011). doi:10.1192/bjp.bp.109.076877

31. Bryson, S. E., Bradley, E. A., Thompson, A. & Wainwright, A. Prevalence of autism among adolescents with intellectual disabilities. *Can. J. Psychiatry* **53**, 449–459 (2008).

32. McGrother, C. W. *et al.* Epilepsy in adults with intellectual disabilities: Prevalence, associations and service implications. *Seizure* **15**, 376–386 (2006).

33. Reilly, C. *et al.* Neurobehavioral comorbidities in children with active epilepsy: A population-based study. *Pediatrics* **133**, e1586–e1593 (2014).

34. Reilly, C. *et al.* Features of autism spectrum disorder (ASD) in childhood epilepsy: A population-based study. *Epilepsy Behav.* **42**, 86–92 (2015).

35. Berg, A. T., Plioplys, S. & Tuchman, R. Risk and correlates of autism spectrum disorder in children with epilepsy: A community-based study. *J. Child Neurol.* **26**, 540–547 (2011).

36. Morgan, V. A., Leonard, H., Bourke, J. & Jablensky, A. Intellectual disability co-occurring with schizophrenia and other psychiatric illness: Population-based study. *Br. J. Psychiatry* (2008). doi:10.1192/bjp.bp.107.044461

37. Lima Caldeira, G., Peça, J. & Carvalho, A. L. New insights on synaptic dysfunction in neuropsychiatric disorders. *Curr. Opin. Neurobiol.* **57**, 62–70 (2019).

38. Torres, F., Barbosa, M. & Maciel, P. Recurrent copy number variations as risk factors for neurodevelopmental disorders: Critical overview and analysis of clinical implications. *J. Med. Genet.* **53**, 73–90 (2015).

39. O'Roak, B. J. *et al.* Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science* **338**, 1619–22 (2012).

40. Turner, E. H., Lee, C., Ng, S. B., Nickerson, D. A. & Shendure, J. Massively parallel exon capture and library-free resequencing across 16 genomes. *Nat. Methods* **6**, 315–316 (2009).

41. Porreca, G. J. *et al.* Multiplex amplification of large sets of human exons. *Nat. Methods* **4**, 931–936

(2007).

42. Krishnakumar, S. *et al.* A comprehensive assay for targeted multiplex amplification of human DNA sequences. *Proc. Natl. Acad. Sci.* (2008). doi:10.1073/pnas.0803240105

43. Hiatt, J. B., Pritchard, C. C., Salipante, S. J., O'Roak, B. J. & Shendure, J. Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. *Genome Res.* **23**, 843–854 (2013).

44. O'Roak, B. J. *et al.* Recurrent de novo mutations implicate novel genes underlying simplex autism risk. *Nat. Commun.* **5**, 5595 (2014).

45. Stessman, H. A., Bernier, R. & Eichler, E. E. A genotype-first approach to defining the subtypes of a complex disease. *Cell* **156**, 872–877 (2014).

46. Stessman, H. A. F., Turner, T. N. & Eichler, E. E. Molecular subtyping and improved treatment of neurodevelopmental disease. *Genome Med.* **8**, 22 (2016).

47. Ronemus, M., Iossifov, I., Levy, D. & Wigler, M. The role of de novo mutations in the genetics of autism spectrum disorders. *Nat. Rev. Genet.* **15**, 133–41 (2014).

48. Bernier, R. *et al.* Disruptive CHD8 mutations define a subtype of autism early in development. *Cell* **158**, 263–276 (2014).

49. Sanders, S. J. *et al.* Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron* **87**, 1215–1233 (2015).

50. Packer, A. Neocortical neurogenesis and the etiology of autism spectrum disorder. *Neurosci. Biobehav. Rev.* **64**, 185–195 (2016).

51. Turner, T. N. *et al.* Proteins linked to autosomal dominant and autosomal recessive disorders harbor characteristic rare missense mutation distribution patterns. *Hum. Mol. Genet.* **24**, 5995–6002 (2015).

52. van Bon, B. W. M. *et al.* Disruptive de novo mutations of DYRK1A lead to a syndromic form of autism and ID. *Mol. Psychiatry* 1–7 (2015). doi:10.1038/mp.2015.5

53. Helsmoortel, C. *et al.* A SWI/SNF-related autism syndrome caused by de novo mutations in ADNP. *Nat. Genet.* **46**, 380–4 (2014).

54. Buxbaum, J. D. DSM-5 and psychiatric genetics - Round hole, meet square peg. *Biol. Psychiatry*

77, 766–768 (2015).

55. Gulsuner, S. *et al.* Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. *Cell* **154**, 518–529 (2013).

56. Hashimoto, R. *et al.* Whole-exome sequencing and neurite outgrowth analysis in autism spectrum disorder. *J.Hum.Genet.* **61**, 1–8 (2015).

57. Helbig, K. L. *et al.* Diagnostic exome sequencing provides a molecular diagnosis for a significant proportion of patients with epilepsy. *Genet. Med.* 1–8 (2016). doi:10.1038/gim.2015.186

58. Homsy, J. *et al.* De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. *Science* **350**, 1262–1266 (2015).

59. Jiang, Y. *et al.* Detection of clinically relevant genetic variants in autism spectrum disorder by whole-genome sequencing. *Am. J. Hum. Genet.* **93**, 249–63 (2013).

60. Kranz, T. M. *et al.* De novo mutations from sporadic schizophrenia cases highlight important signaling genes in an independent sample. *Schizophr. Res.* **166**, 119–124 (2015).

61. Krumm, N. *et al.* Excess of rare, inherited truncating mutations in autism. *Nat. Genet.* **47**, 582–588 (2015).

62. Lee, H., Lin, M. chin A., Kornblum, H. I., Papazian, D. M. & Nelson, S. F. Exome sequencing identifies de novo gain of function missense mutation in KCND2 in identical twins with autism and seizures that slows potassium channel inactivation. *Hum. Mol. Genet.* **23**, 3481–3489 (2014).

63. Lelieveld, S. H. *et al.* Meta-analysis of 2,104 trios provides support for 10 novel candidate genes for intellectual disability. *Nat. Neurosci.* **19**, 1194–1196 (2016).

64. McCarthy, S. E. *et al.* De novo mutations in schizophrenia implicate chromatin remodeling and support a genetic overlap with autism and intellectual disability. *Mol. Psychiatry* **19**, 652–8 (2014).

65. Michaelson, J. J. *et al.* Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell* **151**, 1431–1442 (2012).

66. Rauch, A. *et al.* Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* **380**, 1674–1682 (2012).

67. Tavassoli, T. *et al.* De novo SCN2A splice site mutation in a boy with autism spectrum disorder. *BMC Med. Genet.* **15**, 35 (2014).

68. Veeramah, K. R. *et al.* De novo pathogenic SCN8A mutation identified by whole-genome sequencing of a family quartet affected by infantile epileptic encephalopathy and SUDEP. *Am. J. Hum. Genet.* **90**, 502–510 (2012).

69. Veeramah, K. R. *et al.* Exome sequencing reveals new causal mutations in children with epileptic encephalopathies. *Epilepsia* **54**, 1270–1281 (2013).

70. Yuen, R. K. C. *et al.* Whole-genome sequencing of quartet families with autism spectrum disorder. *Nat. Med.* **21**, 185–91 (2015).

71. Zaidi, S. *et al.* De novo mutations in histone-modifying genes in congenital heart disease. *Nature* **498**, 220–3 (2013).

72. Allen, A. S. *et al.* De novo mutations in epileptic encephalopathies. *Nature* **501**, 217–21 (2013).

73. Barcia, G. *et al.* De novo gain-of-function KCNT1 channel mutations cause malignant migrating partial seizures of infancy. *Nat. Genet.* **44**, 1255–9 (2012).

74. de Ligt, J. *et al.* Diagnostic Exome Sequencing in Persons with Severe Intellectual Disability. *N. Engl. J. Med.* **367**, 1921–1929 (2012).

75. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209–215 (2014).

76. Deciphering Developmental Disorders Study. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **519**, 223–8 (2015).

77. Dimassi, S. *et al.* Whole-exome sequencing improves the diagnosis yield in sporadic infantile spasm syndrome. *Clin. Genet.* 198–204 (2015). doi:10.1111/cge.12636

78. Fromer, M. *et al.* De novo mutations in schizophrenia implicate synaptic networks. *Nature* **506**, 179–184 (2014).

79. Turner, T. N. *et al.* Denovo-Db: a Compendium of Human *De Novo* Variants. *Nucleic Acids Res.* gkw865 (2016). doi:10.1093/nar/gkw865

80. The Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat. Genet.* **46**, 818–825 (2014).

81. Constantino, J. Social Responsiveness Scale (SRS-2). *West. Psychol. Serv.* (2012).

82. Ng, S. B. *et al.* Targeted capture and massicely parallel sequencing of twelve human exomes.

Nature **461**, 272–276 (2010).

83. Fitzgerald, T. W. *et al.* Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **519**, 223–228 (2014).

84. EuroEPINOMICS-RES Consortium, Epilepsy Phenome/Genome Project, and Epi4K Consortium. De novo mutations in synaptic transmission genes including DNM1 cause epileptic encephalopathies. *Am. J. Hum. Genet.* **95**, 360–70 (2014).

85. Turner, T. N. *et al.* Genome sequencing of autism-affected families reveals disruption of putative noncoding regulatory DNA. *Am. J. Hum. Genet.* **98**, 58–74 (2015).

86. Ezkurdia, I. *et al.* Multiple evidence strands suggest that there may be as few as 19000 human protein-coding genes. *Hum. Mol. Genet.* **23**, 5866–5878 (2014).

87. Pirooznia, M. *et al.* SynaptomeDB: An ontology-based knowledgebase for synaptic genes. *Bioinformatics* **28**, 897–899 (2012).

88. Subtil-Rodríguez, A. *et al.* The chromatin remodeller CHD8 is required for E2F-dependent transcription activation of S-phase genes. *Nucleic Acids Res.* **42**, 2185–2196 (2014).

89. Darnell, J. C. *et al.* FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell* **146**, 247–261 (2011).

90. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).

91. McRae, J. F. *et al.* Prevalence and architecture of de novo mutations in developmental disorders. *Nature* (2017). doi:10.1038/nature21062

92. Schuurs-Hoeijmakers, J. H. M. *et al.* Recurrent de novo mutations in PACS1 cause defective cranial-neural-crest migration and define a recognizable intellectual-disability syndrome. *Am. J. Hum. Genet.* **91**, 1122–1127 (2012).

93. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).

94. Lynch, M. Rate, molecular spectrum, and consequences of human mutation. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 961–8 (2010).

95. Banerjee, S., Riordan, M. & Bhat, M. A. Genetic aspects of autism spectrum disorders: insights

from animal models. *Front. Cell. Neurosci.* **8**, 58 (2014).

96. Stessman, H. A. F. *et al.* Targeted sequencing identifies 91 neurodevelopmental-disorder risk genes with autism and developmental-disability biases. *Nat. Genet.* **49**, 515–526 (2017).

97. Boyle, E. a, O'Roak, B. J., Martin, B. K., Kumar, A. & Shendure, J. MIPgen: optimized modeling and design of molecular inversion probes for targeted resequencing. *Bioinformatics* **30**, 2670–2 (2014).

98. Cantsilieris, S., Stessman, H. A., Shendure, J. & Eichler, E. E. Targeted capture and high-throughput sequencing using molecular inversion probes (MIPs). *Methods in Molecular Biology* (2017). doi:10.1007/978-1-4939-6442-0_6

99. Hoischen, A. *et al.* De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. *Nat. Genet.* **42**, 483–485 (2010).

100. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

101. Girirajan, S. *et al.* Refinement and discovery of new hotspots of copy-number variation associated with autism spectrum disorder. *Am. J. Hum. Genet.* **92**, 221–237 (2013).

102. Coe, B. P. *et al.* Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nat. Genet.* **46**, (2014).

103. Moreno-Ramos, O. A., Olivares, A. M., Haider, N. B., De Autismo, L. C. & Lattig, M. C. Whole-exome sequencing in a south American cohort links ALDH1A3, FOXN1 and retinoic acid regulation pathways to autism spectrum disorders. *PLoS One* **10**, 1–13 (2015).

104. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

105. Douville, C. *et al.* CRAVAT: Cancer-related analysis of variants toolkit. *Bioinformatics* **29**, 647–648 (2013).

106. Maheshwari, M. *et al.* PTPN11 mutations in Noonan Syndrome Type I: Detection of recurrent mutations in exons 3 and 13. *Hum. Mutat.* **20**, 298–304 (2002).

107. Myhre, S. A., Ruvalcaba, R. H. A. & Graham, C. B. A new growth deficiency syndrome. *Clin. Genet.* **20**, 1–5 (2008).

108. Le Goff, C. *et al.* Mutations at a single codon in Mad homology 2 domain of SMAD4 cause Myhre

syndrome. *Nat. Genet.* **44**, 85–8 (2012).

109. Landrum, M. J. *et al.* ClinVar: Public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **44**, D862–D868 (2016).

110. Yuan, H., Erreger, K., Dravid, S. M. & Traynelis, S. F. Conserved structural and functional control of N-methyl-D-aspartate receptor gating by transmembrane domain M3. *J. Biol. Chem.* **280**, 29708–29716 (2005).

111. Zuo, J. *et al.* Neurodegeneration in Lurcher mice caused by mutation in delta2 glutamate receptor gene. *Nature* **388**, 769–773 (1997).

112. Coutelier, M. *et al.* GRID2 mutations span from congenital to mild adult-onset cerebellar ataxia. *Neurology* **84**, 1751–1759 (2015).

113. Kohda, K., Wang, Y. & Yuzaki, M. Mutation of a glutamate receptor motif reveals its role in gating and delta2 receptor channel properties. *Nat. Neurosci.* **3**, 315–322 (2000).

114. Taverna, F. *et al.* The Lurcher mutation of an α-amino-3-hydroxy-5-methyl-4- isoxazolepropionic acid receptor subunit enhances potency of glutamate and converts an antagonist to an agonist. *J. Biol. Chem.* **275**, 8475–8479 (2000).

115. Klein, R. M. & Howe, J. R. Effects of the lurcher mutation on GluR1 desensitization and activation kinetics. *J. Neurosci.* **24**, 4941–4951 (2004).

116. Kessels, H. W. & Malinow, R. Synaptic AMPA Receptor Plasticity and Behavior. *Neuron* **61**, 340–350 (2009).

117. Niknafs, N. *et al.* MuPIT interactive: Webserver for mapping variant positions to annotated, interactive 3D structures. *Hum. Genet.* **132**, 1235–1243 (2013).

118. Durand, C. M. *et al.* Mutations in the gene encoding the synaptic scaffolding protein SHANK3 are associated with autism spectrum disorders. *Nat Genet* **39**, 25–27 (2007).

119. Hamdan, F. F. *et al.* De novo syngap1 mutations in nonsyndromic intellectual disability and autism. *Biol. Psychiatry* **69**, 898–901 (2011).

120. Peñagarikano, O. *et al.* Absence of CNTNAP2 leads to epilepsy, neuronal migration abnormalities, and core autism-related deficits. *Cell* **147**, 235–246 (2011).

121. Varoqueaux, F. *et al.* Neuroligins Determine Synapse Maturation and Function. *Neuron* **51**, 741–

754 (2006).

122. Zanjani, H. S. *et al.* Death and survival of heterozygous lurcher purkinje cells in vitro. *Dev. Neurobiol.* **69**, 505–517 (2009).

123. Andrásfalvy, B. K., Smith, M. a, Borchardt, T., Sprengel, R. & Magee, J. C. Impaired regulation of synaptic strength in hippocampal neurons from GluR1-deficient mice. *J. Physiol.* **552**, 35–45 (2003).

124. Wiedholz, L. M. *et al.* Mice lacking the AMPA GluR1 receptor exhibit striatal hyperdopaminergia and 'schizophrenia-related' behaviors. *Mol. Psychiatry* **13**, 631–640 (2008).

125. Samocha, K. E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* **46**, 944–950 (2014).

126. Geisheker, M. R. *et al.* Hotspots of missense mutation identify neurodevelopmental disorder genes and functional domains. *Nat. Neurosci.* **20**, 1043–1051 (2017).

127. Turner, T. N. *et al.* Genomic patterns of de novo mutation in simplex autism. *Cell* **171**, 710-722.e12 (2017).

128. Yuen, R. K. C. *et al.* Genome-wide characteristics of de novo mutations in autism. *NPJ Genomic Med.* (2016). doi:10.1038/npjgenmed.2016.27

129. Yuen, R. K. C. *et al.* Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. *Nat. Neurosci.* (2017). doi:10.1038/nn.4524

130. Halvardson, J. *et al.* Mutations in HECW2 are associated with intellectual disability and epilepsy. *J. Med. Genet.* (2016). doi:10.1136/jmedgenet-2016-103814

131. Wang, T. *et al.* De novo genic mutations among a Chinese autism spectrum disorder cohort. *Nat. Commun.* (2016). doi:10.1038/ncomms13316

132. Hormozdiari, F., Penn, O., Borenstein, E. & Eichler, E. E. The discovery of integrated gene networks for autism and related disorders. 142–154 (2015). doi:10.1101/gr.178855.114.142

133. Warde-Farley, D. *et al.* The GeneMANIA prediction server: Biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.* (2010). doi:10.1093/nar/gkq537

134. Xu, X., Nehorai, A. & Dougherty, J. D. Cell type-specific analysis of human brain transcriptome data to predict alterations in cellular composition. *Syst. Biomed.* (2014). doi:10.4161/sysb.25630

135. Pinto, D. *et al.* Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. *Am. J. Hum. Genet.* **94**, 677–94 (2014).

136. Krumm, N., O'Roak, B. J., Shendure, J. & Eichler, E. E. A de novo convergence of autism genetics and molecular neuroscience. *Trends Neurosci.* **37**, 95–105 (2014).

137. Li, J. *et al.* Integrated systems analysis reveals a molecular network underlying autism spectrum disorders. *Mol. Syst. Biol.* **10**, 774–774 (2014).

138. Takata, A. *et al.* Integrative Analyses of De Novo Mutations Provide Deeper Biological Insights into Autism Spectrum Disorder. *Cell Rep.* **22**, 734–747 (2018).

139. Szklarczyk, D. *et al.* STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43**, D447-52 (2015).

140. Cogné, B. *et al.* Missense Variants in the Histone Acetyltransferase Complex Component Gene TRRAP Cause Autism and Syndromic Intellectual Disability. *Am. J. Hum. Genet.* **104**, 530–541 (2019).

141. Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M. & Ng, P. C. SIFT missense predictions for genomes. *Nat. Protoc.* (2016). doi:10.1038/nprot.2015.123

142. Adzhubei, I. *et al.* PolyPhen-2 : prediction of functional effects of human nsSNPs. *Nat. Methods* (2010). doi:10.1017/CBO9781107415324.004

143. Zoghbi, H. Y. & Bear, M. F. Synaptic dysfunction in neurodevelopmental disorders associated with autism and intellectual disabilities. **4**, a009886--a009886 (2012).

144. Wilfert, A. B., Sulovari, A., Turner, T. N., Coe, B. P. & Eichler, E. E. Recurrent de novo mutations in neurodevelopmental disorders: Properties and clinical implications. *Genome Med.* **9**, 1–16 (2017).

145. Willsey, A. J. *et al.* Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell* **155**, 997 (2013).

146. Herceg, Z. *et al.* Disruption of Trrap causes early embryonic lethality and defects in cell cycle progression. *Nat. Genet.* **29**, 206–211 (2001).

147. McMahon, S. B., Van Buskirk, H. A., Dugan, K. A., Copeland, T. D. & Cole, M. D. The novel ATM-related protein TRRAP is an essential cofactor for the c-Myc and E2F oncoproteins. *Cell* **94**, 363–

374 (1998).

148. Park, J., Kunjibettu, S., McMahon, S. B. & Cole, M. D. The ATM-related domain of TRRAP is required for histone acetyltransferase recruitment and Myc-dependent oncogenesis. *Genes Dev.* **15**, 1619–1624 (2001).

149. Coe, B. P. *et al.* Neurodevelopmental disease genes implicated by de novo mutation and copy number variation morbidity. *Nat. Genet.* **51**, 106–116 (2019).

150. Marie, S. *et al.* AICA-Ribosiduria: A novel, neurologically devastating inborn error of purine biosynthesis caused by mutation of ATIC. *Am. J. Hum. Genet.* **74**, 1276–1281 (2004).

151. Steinberg, S. *et al.* Truncating mutations in RBM12 are associated with psychosis. *Nat. Genet.* **49**, 1251–1254 (2017).

152. Bourdeaut, F. *et al.* ALK germline mutations in patients with neuroblastoma: A rare and weakly penetrant syndrome. *Eur. J. Hum. Genet.* **20**, 291–297 (2012).

153. Chen, Y. *et al.* Oncogenic mutations of ALK kinase in neuroblastoma. *Nature* **455**, 971–974 (2008).

154. Janoueix-Lerosey, I. *et al.* Somatic and germline activating mutations of the ALK kinase receptor in neuroblastoma. *Nature* **455**, 967–970 (2008).

155. Mossé, Y. P. *et al.* Identification of ALK as a major familial neuroblastoma predisposition gene. *Nature* **455**, 930–935 (2008).

156. Amir, R. E. *et al.* Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nat. Genet.* (1999). doi:10.1038/13810

157. Okur, V. *et al.* De novo mutations in CSNK2A1 are associated with neurodevelopmental abnormalities and dysmorphic features. *Hum. Genet.* (2016). doi:10.1007/s00439-016-1661-y

158. Di Meglio, C. *et al.* Epileptic patients with de novo STXBP1 mutations: Key clinical features based on 24 cases. *Epilepsia* (2015). doi:10.1111/epi.13214

159. Chérot, E. *et al.* Using medical exome sequencing to identify the causes of neurodevelopmental disorders: Experience of 2 clinical units and 216 patients. *Clin. Genet.* (2018). doi:10.1111/cge.13102

160. Cordeddu, V. *et al.* Mutation of SHOC2 promotes aberrant protein N-myristoylation and causes

Noonan-like syndrome with loose anagen hair. *Nat. Genet.* (2009). doi:10.1038/ng.425

161. Hamilton, M. J. *et al.* Heterozygous mutations affecting the protein kinase domain of CDK13 cause a syndromic form of developmental delay and intellectual disability. *J. Med. Genet.* (2018). doi:10.1136/jmedgenet-2017-104620

162. Sifrim, A. *et al.* Distinct genetic architectures for syndromic and nonsyndromic congenital heart defects identified by exome sequencing. *Nat. Genet.* (2016). doi:10.1038/ng.3627

163. Rodríguez-Escudero, I. *et al.* A comprehensive functional analysis of PTEN mutations: Implications in tumor- and autism-related syndromes. *Hum. Mol. Genet.* (2011). doi:10.1093/hmg/ddr337

164. Wolff, M. *et al.* Genetic and phenotypic heterogeneity suggest therapeutic implications in SCN2A-related disorders. *Brain* **140**, 1316–1336 (2017).

165. Stamberger, H. *et al.* STXBP1 encephalopathy: A neurodevelopmental disorder including epilepsy. *Neurology* (2016).

166. Amiel, J. *et al.* Mutations in TCF4, encoding a class i basic helix-loop-helix transcription factor, are responsible for Pitt-Hopkins Syndrome, a severe epileptic encephalopathy associated with autonomic dysfunction. *Am. J. Hum. Genet.* (2007). doi:10.1086/515582

167. Houge, G. *et al.* B56δ-related protein phosphatase 2A dysfunction identified in patients with intellectual disability. *J. Clin. Invest.* (2015). doi:10.1172/JCI79860

168. Narumi, Y. *et al.* Molecular and clinical characterization of cardio-facio-cutaneous (CFC) syndrome: Overlapping clinical manifestations with Costello syndrome. *Am. J. Med. Genet.* (2007). doi:10.1002/ajmg.a.31658

169. Palmer, E. E. *et al.* De novo and inherited mutations in the X-linked gene CLCN4 are associated with syndromic intellectual disability and behavior and seizure disorders in males and females. *Mol. Psychiatry* (2018). doi:10.1038/mp.2016.135

170. Lam, W. W. K. *et al.* Novel de novo EEF1A2 missense mutations causing epilepsy and intellectual disability. *Mol. Genet. Genomic Med.* (2016). doi:10.1002/mgg3.219

171. Kortüm, F. *et al.* Mutations in KCNH1 and ATP6V1B2 cause Zimmermann-Laband syndrome. *Nat. Genet.* (2015). doi:10.1038/ng.3282

172. Weckhuysen, S. *et al.* Extending the KCNQ2 encephalopathy spectrum: Clinical and neuroimaging findings in 17 patients. *Neurology* (2013). doi:10.1212/01.wnl.0000435296.72400.a1

173. Rodriguez-Viciana, P. *et al.* Germline mutations in genes within the MAPK pathway cause cardio-facio-cutaneous syndrome. *Science* (2006). doi:10.1126/science.1124642

174. Bourdon, V. *et al.* A detailed analysis of the MECP2 gene: Prevalence of recurrent mutations and gross DNA rearrangements in Rett syndrome patients. *Hum. Genet.* (2001). doi:10.1007/s004390000422

175. Heilstedt, H. A., Shahbazian, M. D. & Lee, B. Infantile hypotonia as a presentation of Rett syndrome. *Am. J. Med. Genet.* (2002). doi:10.1002/ajmg.10633

176. Dedek, K., Fusco, L., Teloy, N. & Steinlein, O. K. Neonatal convulsions and epileptic encephalopathy in an Italian family with a missense mutation in the fifth transmembrane region of KCNQ2. *Epilepsy Res.* (2003). doi:10.1016/S0920-1211(03)00037-8

177. Simons, C. *et al.* Mutations in the voltage-gated potassium channel gene KCNH1 cause Temple-Baraitser syndrome and epilepsy. *Nat. Genet.* (2015). doi:10.1038/ng.3153

178. Saitsu, H. *et al.* De novo mutations in the gene encoding STXBP1 (MUNC18-1) cause early infantile epileptic encephalopathy. *Nat. Genet.* (2008). doi:10.1038/ng.150

179. Hamdan, F. F. *et al.* High rate of recurrent de novo mutations in developmental and epileptic encephalopathies. *Am. J. Hum. Genet.* **101**, 664–685 (2017).

180. Gregor, A. *et al.* De novo mutations in the genome organizer CTCF cause intellectual disability. *Am. J. Hum. Genet.* (2013). doi:10.1016/j.ajhg.2013.05.007

181. Vissers, L. E. L. M. *et al.* Mutations in a new member of the chromodomain gene family cause CHARGE syndrome. *Nat. Genet.* (2004). doi:10.1038/ng1407

182. Sergouniotis, P. I. *et al.* Biallelic variants in TTLL5, encoding a tubulin glutamylase, cause retinal dystrophy. *Am. J. Hum. Genet.* (2014). doi:10.1016/j.ajhg.2014.04.003

183. Zweier, M. *et al.* Mutations in MEF2C from the 5q14.3q15 microdeletion syndrome region are a frequent cause of severe mental retardation and diminish MECP2 and CDKL5 expression. *Hum. Mutat.* (2010). doi:10.1002/humu.21253

184. Najmabadi, H. *et al.* Deep sequencing reveals 50 novel genes for recessive cognitive disorders.

Nature (2011). doi:10.1038/nature10423

185. Fukai, R. *et al.* De novo KCNH1 mutations in four patients with syndromic developmental delay, hypotonia and seizures. *J. Hum. Genet.* (2016). doi:10.1038/jhg.2016.1

186. Miceli, F. *et al.* Early-onset epileptic encephalopathy caused by gain-of-function mutations in the voltage sensor of Kv7.2 and Kv7.3 potassium channel subunits. *J. Neurosci.* (2015). doi:10.1523/jneurosci.4423-14.2015

187. Liman, E. R., Hess, P., Weaver, F. & Koren, G. Voltage-sensing residues in the S4 region of a mammalian K + channel. *Nature* **353**, 752–756 (1991).

188. Lal, D. *et al.* Gene family information facilitates variant interpretation and identification of disease-associated genes. *bioRxiv* (2017). doi:10.1101/159780

189. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015).

190. Jarvik, E. R. *et al.* Actionable exomic incidental findings in 6503 participants: challenges of variant classification. *Genome Res.* **25**, 305–315 (2015).

191. Berkel, S. *et al.* Mutations in the SHANK2 synaptic scaffolding gene in autism spectrum disorder and mental retardation. *Nat. Genet.* **42**, 489–491 (2010).

192. Gilman, S. R. *et al.* Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. *Neuron* **70**, 898–907 (2011).

193. Sanders, S. J. *et al.* Progress in Understanding and Treating SCN2A-Mediated Disorders. *Trends Neurosci.* **41**, 442–456 (2018).

194. Stessman, H. A. F. *et al.* Disruption of POGZ Is Associated with Intellectual Disability and Autism Spectrum Disorders. *Am. J. Hum. Genet.* **98**, 541–552 (2016).

195. Küry, S. *et al.* De novo mutations in protein kinase genes CAMK2A and CAMK2B cause intellectual disability. *Am. J. Hum. Genet.* **101**, 768–788 (2017).

196. Cheng, H. *et al.* Truncating variants in NAA15 are associated with variable levels of intellectual disability, autism spectrum disorder, and congenital anomalies. *Am. J. Hum. Genet.* **102**, 985–994 (2018).

197. Van Dijck, A. *et al.* Clinical presentation of a complex neurodevelopmental disorder caused by mutations in ADNP. *Biol. Psychiatry* **85**, 287–297 (2019).

198. McMahon, S. B., Wood, M. A. & Cole, M. D. The essential cofactor TRRAP recruits the histone acetyltransferase hGCN5 to c-Myc. *Mol. Cell. Biol.* **20**, 556–62 (2000).

199. Feliciano, P. *et al.* SPARK: A US cohort of 50,000 families to accelerate autism research. *Neuron* **97**, 488–493 (2018).

200. Fowler, D. M. *et al.* High-resolution mapping of protein sequence-function relationships. *Nat. Methods* **7**, 741–6 (2010).

201. Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat. Methods* **11**, 801–807 (2014).

202. Weile, J. *et al.* A framework for exhaustively mapping functional missense variants. *Mol. Syst. Biol.* **13**, 957 (2017).

203. Starita, L. M. *et al.* Massively parallel functional analysis of BRCA1 RING domain variants. *Genetics* **200**, 413–422 (2015).

204. Rubaiy, H. N. A short guide to electrophysiology and ion channels. *J. Pharm Pharm Sci* **20**, 48–67 (2017).

205. Rocha-Martins, M., Cavalheiro, G. R., Matos-Rodrigues, G. E. & Martins, R. A. P. From gene targeting to genome editing: Transgenic animals applications and beyond. *An. Acad. Bras. Cienc.* **87**, 1323–1348 (2015).

206. Shrock, E. & Güell, M. CRISPR in animals and animal models. *Progress in Molecular Biology and Translational Science* (2017). doi:10.1016/bs.pmbts.2017.07.010

207. Hruscha, A. *et al.* Efficient CRISPR/Cas9 genome editing with low off-target effects in zebrafish. *Development* **140**, 4982–4987 (2013).

208. Shinoda, Y., Sadakata, T. & Furuichi, T. Animal models of autism spectrum disorder (ASD): a synaptic-level approach to autistic-like behavior in mice. *Exp. Anim.* **62**, 71–8 (2013).